

14 May 2010 | \$10

Science

Tuberculosis
& Malaria



AAAS



Barry R. Bloom is
Harvard University
Distinguished Service
Professor of Public
Health and former dean
of the Harvard School
of Public Health.
E-mail: barry_bloom@harvard.edu

Support for Global Health

AS MORE NATIONS STRUGGLE WITH STRESSED ECONOMIES, AID TO THE DEVELOPING WORLD BECOMES increasingly vulnerable to governments' budgetary cuts. The industrialized world is recognizing that coordinating global development assistance is the most efficient way to maximize effectiveness and minimize duplication. Earlier this year, the United States, the largest funder of global health assistance, announced that it seeks to expand multilateral efforts to address the major health problems of developing countries. One of the triumphs of multilateral cooperation has been the Global Fund to Fight AIDS, Tuberculosis and Malaria, a program that has saved millions of lives in developing countries. That is why the \$50 million reduction in funding for the Global Fund requested by the U.S. government for fiscal year 2011, in the face of increased requests for expanded coverage by those countries, would be a major setback.

The Global Fund was created in 2002 as an innovative multilateral agency to mobilize aid at a global scale. As both a funding and a knowledge agency, it operates under the principle of country ownership, in which a country's coordinating mechanism (with representatives from government, nonprofit organizations, the private sector, and civil society) takes ownership of the project, formulating and implementing programs on the ground. A portion of the Global Fund is also allocated to strengthening a country's burdened health system more generally, thus minimizing the diversion of limited resources from other needed programs. Thus far, it has supported comprehensive prevention, treatment, and care programs in 144 countries through more than \$19 billion of donor financing. Annually, it provides about 57% of all international financing for tuberculosis treatment, 60% for malaria, and 23% for HIV. Before funding is renewed, each project is rigorously reviewed and evaluated.

Approximately 5 million people would have died of AIDS, tuberculosis, or malaria over the past 5 years if not for the interventions supported by the Global Fund. It has enabled 2.5 million people to receive antiretroviral drug therapy for HIV. In addition, 6 million people have been provided with effective tuberculosis treatment, and 104 million insecticide-treated bed nets have been distributed to protect families from malaria. Funding has also gone toward care for orphans; preventing transmission of HIV from mother to child; and malaria treatment, with a major thrust to use artemisinin-based combination therapy to prevent resistance to what is currently the most useful drug for malaria.

For every \$1 the U.S. government contributes, the Global Fund leverages \$2 in contributions from other donor governments. Decreasing support for this fund would thus be inconsistent with the proclaimed new multilateral objectives of the Obama Administration. The U.S. International Affairs budget, requested for 2011 at \$58.5 billion, represents only 1.5%

of the federal budget, with health constituting \$10.43 billion of that amount. Yet last month, the Senate Budget Committee announced a cut of \$4 billion, or 7%. Given the success of the President's Emergency Program for AIDS Relief (PEPFAR) and Global Fund programs, cuts in the health component could have a tragic impact on millions of lives. Bilateral assistance such as PEPFAR, now in 30 countries, has value to the United States in both political and humanitarian terms. But the United States cannot itself be doctor to the world. Increased multilateral support through the Global Fund would greatly expand the U.S. contribution, ensure efficiency, and enable shared responsibility with other donors.

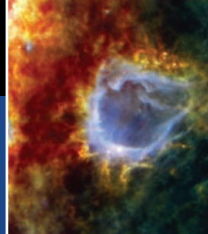
By announcing its intent to expand multilateral efforts, the United States has committed itself to maintain leadership in international development, which the White House states to be as critical to national security as defense and diplomacy. This commitment should include vigorous funding for global health, including the Global Fund. In a turbulent world with huge disparities of wealth, health, and education, all nations must support partnerships that effectively direct the limited resources available toward helping the world's poorest live a better life.

— Barry R. Bloom

10.1126/science.1189538



Online
sciencemag.org
 Listen to the
author online.



Herschel's first results

802



Sweden enters the genomics era

805

FUSION

ITER Cost Estimates Leave Europe Struggling to Find Ways to Pay

For the past 2 years, officials involved in the ITER fusion reactor project—poised to begin construction in France—have muttered darkly about huge increases in its estimated cost. Now, finally, we have some hard figures: Last week, the European Commission, the executive of the European Union, released a memo announcing that Europe's contribution to building ITER (45% of the total) would amount to €7.2 billion, 2.7 times the original estimate. Crucially, E.U. fusion funding up to the end of 2013 is short by €1.4 billion.

Senior fusion researchers in Europe who spoke with *Science* were astonished that the commission



Waiting to start? June's ITER council is expected to appoint Osamu Motojima director general and give a final green light for construction at Cadarache (above).



proposed no solution to the problem apart from asking E.U. member states for more money. With Europe's wider economy in turmoil, this funding crisis must be resolved before the mid-June ITER council meeting when the other partners expect to give the project's design, cost, and schedule final approval. (The June meeting is also expected to approve the appointment of a new director general for ITER, reported to be Osamu Motojima of the National Institute for Fusion Science in Toki City, Japan.)

ITER aims to show that fusing hydrogen nuclei inside a huge and complex reactor called a tokamak can produce enough excess energy to be a viable energy source. When the

seven partners in the project—China, the E.U., India, Japan, South Korea, Russia, and the United States—signed the project into existence in 2006, they based their cost estimates on the design that had been adopted in 2001. Soon after the new ITER staff began work at the chosen site of Cadarache, it became clear that those estimates were far too low.

Under ITER's unusual funding arrangement, each partner must manufacture and deliver a share of the reactor components to the site. How much it spends making them

is its own concern. Each partner apart from the E.U. is responsible for 9% of the reactor's "value"; the E.U., as host, supplies 45%. All of the partners got a nasty shock when they looked in detail at what they had agreed to make. The United States, for example, which had estimated its share would be \$1 billion, is now expecting a bill for up to \$2.2 billion.

For the E.U., with its proportionately larger share, the inflated cost is causing serious problems. In documents accompanying last week's memo, the commission blames the increases on a number of factors, including updates to the reactor design, the cost of setting up the organization to manage its procurement, the added complexity of hav-

ing seven partners (there were only three in 2001), and extra resources for quality control. When the scale of the increases became apparent, the E.U. pushed for the construction schedule to be stretched out and called for inquiries into cost estimation and management structures (*Science*, 27 June 2008, p. 1707). Europe's stance has frustrated the other ITER partners. "Even if the budgetary situation is severe, it's important that every country properly fulfill its obligations under this international agreement," says Yoshiyuki Chihara, director of the International Nuclear and Fusion Energy Affairs Division at Japan's Ministry of Education.

Last week's memo spelled out the budget tangle the commission has got itself into. The E.U. multiyear budget that supports ITER runs until the end of 2013. For the years 2012 and 2013, there is only about €700 million in the pot, but according to the memo the commission needs €2.1 billion during that time for procurements early in ITER construction. Commission officials looked into two possible ways of filling that gap. The first was a loan from the European Investment Bank, an E.U. institution that lends to development projects within Europe and, increasingly, large-scale R&D efforts. But the memo dismisses this option because it says there is no identifiable income stream to repay the loan—a view that one fusion researcher called "totally silly." The second option is to transfer funds from other budget areas, but this too is rejected. "The budget is very tight; ... there is now hardly any space for additional shifting of money between budget lines," says Herbert Reul, a member of the European Parliament and chair of its industry, research, and energy committee.

In summary, the commission says "a sustainable solution will require a clear financial commitment for the life of the project and a commitment to finance any overruns outside that framework, by member states." Researchers who spoke with *Science* say the commission got itself into this mess and now wants the E.U. member states to bail it out. "Europe is hugely obligated. It went out on a limb to lead the project. ... Having taken it on, it has to find a way," says one who asked not to be identified. Says another: "If the E.U. doesn't do everything to regain the confidence of the other partners, it is lost."

—DANIEL CLERY

With additional reporting by Dennis Normile.

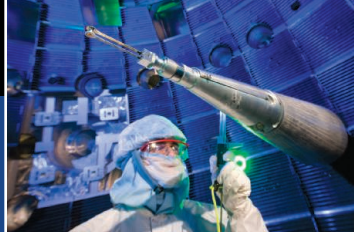
CREDITS (LEFT TO RIGHT): © AGENCE ITER FRANCE; NATIONAL INSTITUTION FOR FUSION SCIENCE

Downloaded from www.sciencemag.org on May 13, 2010



The laser at 50:
New frontiers

806



AIDS RESEARCH

Bioethicists Assail a Celebrated TB/HIV Treatment Trial

A heralded clinical study in South Africa that assessed the best treatment strategy for people infected with HIV who are receiving drugs for tuberculosis should never have been done, argue two bioethicists in an online posting published 5 May by the Hastings Center, a bioethics research institute based in Garrison, New York.

The study, *Starting Antiretroviral Therapy at Three Points in Tuberculosis* (SAPIT), began in June 2005 and was stopped in September 2008 after an interim analysis found that people who delayed starting anti-HIV drugs until they completed a 6-to-8-month course of anti-TB medication had twice the risk of death. The prominent research group that ran the 642-person study immediately reported the findings, which led the World Health Organization (WHO) to strengthen its guidelines; a full report appeared in the 25 February 2010 issue of *The New England Journal of Medicine* (NEJM).

Bioethicists Sean Philpott and Udo Schüklenk branded the study “deeply flawed,” arguing that it violated the Declaration of Helsinki as it “caused foreseeable harms and preventable deaths for a substantial number

of impoverished and poorly educated South African trial participants.” They further lambasted the South African ethics committee that approved SAPIT, contending that its failure “to recognize the clinical, ethical, and legal deficiencies in this study is shameful.” They also noted that U.S. co-authors shared “the academic glory” but did not submit the study protocol to their institutions for ethical review. Philpott is based at Union Graduate College in Schenectady, New York, and Schüklenk, who works at Queen’s University in Kingston, Canada, is co-editor of *Bioethics*.

The head of SAPIT, Salim Abdool Karim of the University of KwaZulu-Natal in Durban, says he is stunned by the criticisms. “When I looked at it, my head fell down,” says Karim, an epidemiologist who also directs the Centre for the AIDS Program of Research in South Africa, which ran the study. At the trial’s start, says Karim, many clinicians worried about the dangers of simultaneously treating HIV and TB for three main reasons: TB antibiotics can weaken the effectiveness of antiretroviral drugs; ARVs can be toxic, leading people to stop taking all medicines; and mixing the two treatments increases the risk

of a life-threatening disease called immune reconstitution syndrome. “The only reason we got involved in the study is because it was so obvious to us that we had no idea what we were doing with these patients,” says Karim, whose country has more HIV/TB coinfecting people than any in the world. “When you put them on ARVs, they died. When you didn’t, they died. We were at sea.”

Karim notes that the 2003 WHO guidelines, *Scaling Up Antiretroviral Therapy in Resource-Limited Settings*, explicitly acknowledges the confusion. “The optimal time to initiate ART in patients with TB is not known,” the guidelines state. In particular, complications from treating TB and HIV simultaneously are most acute during the intensive initial phase of TB drugs, which lasts 2 to 3 months. Thus the guidelines suggest that “deferring the start of ART may be reasonable in a variety of clinical scenarios.” They advise, however, that patients with advanced disease—defined as fewer than 200 CD4 white blood cells per milliliter of blood—“be started 2 weeks to 2 months after the start of TB therapy.”

The critics’ main complaint concerns these severely ill people. In the SAPIT study, 213 people with an average CD4 count of 140 were randomized to delay ARV treatment until completing TB therapy, and 429 others with an average of 150 CD4s started ARVs either within 4 weeks of TB treatment or within 4 weeks of ending the initial intensive phase. The delayed arm had 27 deaths versus 25 in the much larger “integrated” treatment group. All but eight deaths occurred in people with fewer than 200 CD4s, and given that some people died who received ARVs early, the critics calculate that at least 10 “preventable” deaths occurred in the delayed arm. “It is true that the question of when to initiate ARV in TB patients was an open question,” says Schüklenk. “To move from there to the proposition that it was acceptable to delay ARVs in patients even with very low CD4 counts by up to 9 months is false.”

Karim points out that one of the most surprising findings of the study is that there was no significant mortality difference between the integrated and delayed arms until well ▶



CREDIT: CAPRISA

Rewriting history? Salim Abdool Karim says the criticism of the SAPIT study he led discounts the widespread confusion that existed 5 years ago about when to start HIV treatment in patients receiving TB drugs.

ScienceNOW

after people stopped TB treatment. This, he says, underscores why so much confusion existed before the SAPIT results. “The deaths are really occurring past 10 months,” he says, noting that TB doctors typically have stopped seeing the patients by then and thus would never know whether early use of ARVs was critical.

Karim further notes that the SAPIT study design, in keeping with WHO guidelines, allowed “highly trained physicians” to start ARVs based on their judgment about an individual’s needs. But Philpott and Schüklenk don’t believe that occurred. “We really don’t think those individuals randomized to the delayed arm were receiving the individualized care,” says Philpott.

Opinions of clinicians and public health officials who were not involved with the study vary widely. “I think most medical observers familiar with current HIV/AIDS research practice in Africa, with the rapid evolution of WHO guidelines, and with programmatic work on HIV/AIDS in low- and middle-income countries, would consider these criticisms unjustified,” says Kevin De Cock, the former director of WHO’s HIV/AIDS program and now head of the Center for Global Health at the U.S. Centers for Disease Control and Prevention. “The Hastings piece is written very much from the perspective of what we know now, not how we viewed practice in 2003–4 when discussions about how to handle HIV-associated tuberculosis led to studies such as this.”

Some South African clinicians who treat coinfecting HIV/TB patients agree with the critics. “I would never have allowed one of my patients or a family member to enter this study—it was just too dangerous,” says Francois Venter of the University of the Witwatersrand in Johannesburg. Venter stresses that coinfecting patients with fewer than 200 CD4s can get sick and die in a day or two. Gary Maartens, a clinical pharmacologist at the University of Cape Town, adds that had the study limited enrollment to people with between 200 and 500 CD4 cells, it still would have answered an important question about timing of ARVs. “We did not need a clinical trial to tell us that deferring lifesaving therapy for 8 months in patients with advanced disease results in higher mortality,” says Maartens.

Still other HIV/TB specialists say the issue of including people with CD4 levels under 200 is more complex. “It’s all too easy to say this shouldn’t have been done,” says William

Burman, who directs the HIV clinic at Denver Public Health in Colorado and specializes in designing HIV/TB clinical trials. Although Burman says he “had concerns” about the SAPIT study’s design and that the results from people with the lower CD4 counts did not surprise him, he finds the Hastings critique “overly negative.” Despite WHO guidelines, many TB clinicians in poor, high-burden countries had little experience with ARVs when the study began and were reluctant to treat patients with them. “SAPIT has clarified that matter and will surely save many, many thousands of lives,” he says.

“The SAPIT study caused foreseeable harms and preventable deaths for a substantial number of impoverished and poorly educated South African trial participants.”

—SEAN PHILPOTT AND UDO SCHÜKLENK,
BIOETHICS FORUM, HASTINGS CENTER WEB SITE

As for the criticism of the ethical review process, two U.S.-based co-authors of the *NEJM* paper, Gerald Friedland of Yale University and Waafa El-Sadr of Columbia University, say they contributed only to the writing of the report. Neither helped design the protocol, saw patients, or provided financial assistance. The U.S. Department of Health and Human Services Office for Human Research Protections does not require institutions to conduct reviews of protocols for co-authors under these conditions. Philpott counters that in the paper, Karim also notes that he has an affiliation with Columbia. Karim says his part-time, adjunct appointment mainly involves teaching and supervision of postgraduate students.

Friedland and El-Sadr, both highly respected HIV/AIDS researchers, strongly object to the SAPIT criticism. “I’m flabbergasted,” says El-Sadr. “This will be cited as one of the landmark studies in HIV/TB.” Friedland adds that the South African ethical review was done by people with proper credentials, and he says WHO’s 2003 guidelines were based just on observational data, a much lower standard of proof than a randomized, controlled clinical trial. “Now the guidelines can be very clear and firm,” he says.

The debate likely will continue over the next few months. In July, *Bioethics* plans to run an extended version of the critique that appeared on the Hastings Center Web site. Many researchers also expect that *NEJM* will run letters to the editor and a reply in the next few weeks.

—JON COHEN

From *Science’s* Online Daily News Site

How Jupiter Got Its Stripes

Jupiter and other gaseous planets are covered from pole to pole with stripes. But astronomers aren’t exactly sure how they got there. Now a team of physicists reports that Jupiter’s stripes may be produced in part by tides, a result of the gravitational tugging of its 60-odd moons. To back up that idea, they’ve performed an experiment to show that tides can produce just such stripes, at least in a very simple mockup of a gaseous planet. <http://bit.ly/jupiter-stripes>

Bonding With Offspring Grows New Neurons in the Mouse Brain

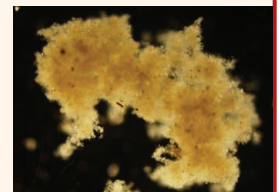
For a father to truly bond with his children, he needs to grow some new gray matter. At least that seems to be the case in mice. A new study shows that when a mouse father nuzzles his pups, he develops new neurons that help him remember—and protect—those offspring later in life. The results suggest that in mice, and perhaps in humans, young babies and dads bond biologically in ways that can last a lifetime. <http://bit.ly/mouse-bonding>

X-ray Vision, Without the Radiation

X-ray-like imaging without the harmful radiation and cell phones with more bandwidth are closer to reality now that researchers have developed a novel type of lens that works with terahertz frequencies. The new lens is a metamaterial, an artificial material with a structure made from many tiny parts, and it could drastically expand what lenses can do. <http://bit.ly/terahertz>

Kon-Tiki, Bacteria Style

New research reveals that, just like a rat clinging to a piece of floating wood after a flood, pathogenic microorganisms can set up shop aboard drifting bits of fish feces and other debris and ride them to far-flung destinations. Understanding more about the process should help public health officials develop better strategies to fight waterborne diseases and seafood contamination. <http://bit.ly/kon-tiki>



Read the full postings, comments, and more at news.sciencemag.org/sciencenow.

ASTROPHYSICS

Herschel's First Images Spark Star-Formation Debate

When the Herschel telescope was launched into space last May, it promised new insights into our universe, detecting far-infrared and submillimeter wavelengths of radiation that have eluded previous missions (*Science*, 1 May 2009, p. 584). A year later, the European Space Agency instrument has passed its commissioning phase and is starting to deliver on that promise, scientists reported at the Herschel First Results Symposium in the Netherlands last week. "Some instruments are working even better than anticipated from the ground tests," says astrophysicist Annie Zavagno of the Laboratoire d'Astrophysique de Marseille, France.

Herschel is the largest infrared telescope ever to be put into space, with a 3.5-meter mirror that's four times larger than any space-based predecessor. There are larger infrared telescopes in ground-based observatories, but Earth's atmosphere absorbs many infrared wavelengths that are vital for studying the births of stars. "The earliest stages of star formation are hidden from our view but can be revealed in the far-infrared

and millimeter domain. This is why Herschel is so important for us and will open a new window on our understanding of star formation," says Zavagno.

One of the results reported last week is already provoking debate. Astronomers had thought that galaxies have been forming stars at about the same rate for the past 3 billion years. However, Herschel observations of nearby galaxies suggest this isn't the case. The telescope detected strong emissions of infrared radiation in those galaxies, which indicates furious star formation, says Steve Eales of Cardiff University in the United Kingdom, a joint principal investigator on the Herschel ATLAS project. Eales and his colleagues conclude that in the past many galaxies were forming stars at 10 to 15 times the rate that we see in our galaxy today.

However, cosmologist Carlton Baugh of Durham University in the U.K. calls this interpretation "controversial." Baugh and his colleagues, including Cedric Lacey, also at Durham University, have been using computer simulations of how galaxies form

to make detailed predictions about what Herschel should find. In a paper published online in March in the *Monthly Notices of the Royal Astronomical Society*, they argued that determining star-formation rates from Herschel observations would be complicated because the telescope can only detect infrared radiation powered by massive stars. To get an overall star-formation rate, scientists must estimate the number of low-mass stars by extrapolating from a relationship that describes the distribution of mass within a newly formed group of stars.

This relationship is typically assumed to be universal for all stellar regions, but Baugh and his colleagues argue that the merger of two galaxies can trigger star formation in which a different mix of stellar masses is produced, with a higher-than-usual proportion of high-mass stars. "Star formation like this can produce more emissions at the long wavelengths probed by Herschel for a modest amount of star formation," says Baugh.

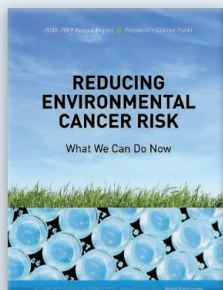
Astronomers may be at odds over the explanation of Herschel's observations of strong infrared emissions from nearby galaxies, but they are in agreement on a second observation: They can't explain it at all. Herschel spied an "impossible" star in the early stages of formation. It has a mass eight to 10 times that of our sun; such a massive

CANCER

Panel Finds Environmental Risks Neglected

A flap has erupted among health experts over a report last week from a presidential advisory group that concludes environmental pollutants are an underappreciated cause of cancer that is inflicting "grievous harm" on Americans. The report, from the President's Cancer Panel, has drawn fire both from an industry-leaning group and from the mainstream American Cancer Society (ACS) for distorting environmental cancer risks.

One of the report's authors, Margaret Kripke, an immunologist and professor emerita at MD Anderson Cancer Center in Houston, Texas, acknowledges that the text is "deliberately thought-provoking" but defends its scientific rigor. "Part of the issue here is to stimulate a call to action, and you don't get that by having a bland" report, Kripke says. She is one of the two members of the authoring panel; the other is LaSalle Leffall, a surgeon at Howard University. A third member, cancer activist Lance Armstrong, who did not contribute



Thought provoking. Co-author Margaret Kripke did not want a report on cancer risks to be bland.

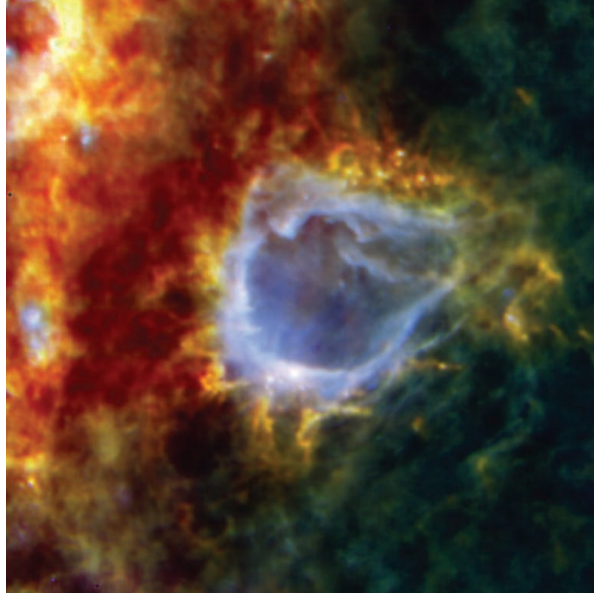
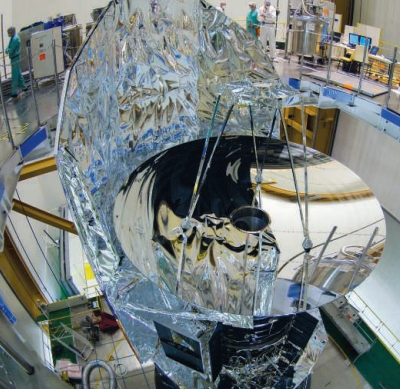
to the report, left the panel in 2008 and was not replaced.

The 240-page report, called *Reducing Environmental Cancer Risk: What We Can Do Now*, states that "the true burden of environmentally induced cancer has been grossly underestimated" and that Americans "are bombarded continually with myriad combinations" of carcinogenic pollutants. It reviews exposures from a range of sources,

including hot-button ones such as the plastics ingredient bisphenol-A, medical x-rays, and even cell phones. Among other recommendations, it urges people to use headsets with cell phones and recommends that Congress and the president adopt the "precautionary approach," which would require manufacturers to demonstrate the safety of chemicals before distributing them on the market.

The American Council on Science and Health, a nonprofit organization with industry support, wrote that "the report practically plagiarizes the work of anti-chemical activist groups" including the Environmental Working Group (EWG), an environmental advocacy organization in Washington, D.C. EWG's Richard Wiles was one of 45 witnesses—including many government and academic scientists—who spoke before the presidential panel in 2008 and early 2009.

Among those who object to the report's conclusions is ACS's emerita chief epidemiologist, Michael Thun. In a response last week, Thun wrote that ACS had issued a report on environmental risks last year that raised similar concerns about untested chem-



star will emit intense ultraviolet radiation that should blast away the cloud of gas that is feeding its growth. Yet the star spotted by Herschel, found within the star-forming cloud RCW 120 in our galaxy, is still surrounded by an envelope of gas and is therefore expected to get even bigger. Astronomers have seen much more massive “impossible” stars before—the star Eta Carinae is thought to be more than 100 times as massive as the sun—but this is the first time that one has been observed at such an early stage of its formation.

Aside from studying the birth of stars, astronomers also hope to use Herschel to analyze the chemical composition of comets and planetary bodies and to examine the molecular chemistry of the universe. The

Seeing is believing. Herschel (left) has spotted an “impossible” massive star, shown in the above image as a white ball near the edge of the blue bubble of gas that depicts the star-forming region RCW 120.

mission team is reporting some success with the latter, as Herschel’s HIFI spectrometer has detected ionized water vapor, the first time that this ion has been detected outside the solar system.

A special issue of *Astronomy & Astrophysics* dedicated to Herschel’s first results is due out this summer, but scientists predict that the instrument will fill numerous pages of journals for many more years to come.

—SARAH REED

icals, the risks to children, and hazards of medical imaging. But Thun finds last week’s report “unbalanced” because of its “dismissal” of cancer prevention focused on lifestyle factors such as smoking and obesity that make a much larger contribution to cancer incidence. Thun also disputed the report’s claim that estimates of pollution-induced cancer are “woefully out of date,” asserting that this statement “reflects one side of a scientific debate.”

That debate centers on the low estimates in 1981 from University of Oxford epidemiologists Richard Doll and Richard Peto, who found that occupational exposures and pollution cause about 6% of cancers. Boston University epidemiologist Richard Clapp, who testified before the presidential panel, acknowledges that he and a few other academics have contested the Doll-Peto figures and have been at odds with ACS on the issue for years.

Peto, for his part, says the report’s criticisms of his and Doll’s methodology are familiar and “factually untrue.” He says, “I don’t think there’s any good reason to

revise our estimates upward.”

The President’s Cancer Panel, created by the 1971 Cancer Act, is independent, although staffed by a small office at the National Cancer Institute. A science writer compiles testimony and the panelists’ own input and writes the report, says the panel’s special assistant, Jennifer Burt. The report does not go through peer review, but staff fact-check the information, she said. She defends the report as “very measured.”

Kripke says she’s “disappointed” by ACS’s criticisms. She says they’re ignoring previous reports from the same panel emphasizing the importance of lifestyle factors. As for the Doll and Peto estimates, she says the view that they need updating is “an opinion.”

Activists are hoping the report will spur efforts in Congress to make U.S. chemical safety regulations more consistent with the precautionary principle. But it’s not clear whether the proposal will gain much momentum. Its immediate destiny is to join a shelf full of previous annual reports from this panel.

—JOCELYN KAISER

ScienceInsider

From the Science Policy Blog



Methane-trapping ice, which has frustrated the first attempt to contain oil gushing off the shore of Louisiana, may have been a root cause of the blowout in the first place. So says University of California, Berkeley, professor Robert Bea, who has extensive access to company documents on the incident. There were signs that drillers did encounter hydrates: About a month before the blowout, a “kick” of gas pressure hit the well hard enough that the platform was shut down. <http://bit.ly/cHFwsa>

Biomedical science will lose a longtime champion with the retirement this year of Representative David Obey (D-WI), the chair of the House Appropriations Committee. Obey, 71, announced that he will not run for reelection this fall after serving 21 terms in Congress. <http://bit.ly/ccUrg>

National Science Foundation **Director Arden Bement offered a sober assessment** of likely congressional action on the 2011 budget after members of the National Science Board suggested that prospects looked rosy. “You should not be surprised if we don’t get the president’s request,” Bement said about the \$552 million (8%) increase for NSF that President Barack Obama proposed in February as part of his overall \$3.6 trillion budget. Bement also said “I won’t be surprised to see us operating under a continuing resolution” until well after the November congressional elections, with funding frozen at current levels. <http://bit.ly/9BeDOq>

The most ambitious U.S. effort to assess environmental change on a continental scale has won final approval from the oversight body of the National Science Foundation. More than a decade in the making, the \$434 million National Ecological Observatory Network (NEON) will establish 20 permanent monitoring stations to collect climate, environmental, and biological data on an ongoing basis. NEON will also include 40 temporary terrestrial sites and 46 aquatic sites. <http://bit.ly/92s93d>

For the full postings and more, go to news.sciencemag.org/scienceinsider.

SCIENTIFIC FACILITIES

Sweden Bets on New Lab to Spruce Up Its Bioscience Future

The tree bears the name of a neighbor—Norway spruce (*Picea abies*)—but this conifer is central to Sweden's fiscal health. It feeds the nation's vigorous timber industries and by some estimates is economically the country's most important species. So it wasn't surprising when the Sweden-based Knut and Alice Wallenberg Foundation last year announced it would provide about \$10 million for the sequencing of the Norway spruce's genome. But what is unexpected—or at least it would have been a year ago—is that this sequencing will largely happen in Sweden rather than being farmed out to other countries. Next week, officials in Stockholm will inaugurate a new building with cutting-edge DNA sequencing machines that by 2013 should produce a rough draft of the genome of the “Christmas” tree.

For Sweden, this building represents far more than a single research project. It's the most visible evidence to date of the Science for Life Laboratory (SciLifeLab), an unusual initiative spanning multiple Swedish research organizations and two sites, one in Stockholm and the other in nearby Uppsala. Rather than spread its money to as many scientists as possible, as it has traditionally done, the Swedish government will spend more than \$75 million setting up SciLifeLab with the hope that it will become a technology-driven national life sciences center, comparable to organizations such as the Broad Institute in Cambridge, Massachusetts. “We want to make sure [the latest] technologies are available to all of Sweden. It's a deliberate decision to do things differently,” says geneticist Kerstin Lindblad-Toh, who is the director of SciLifeLab's Uppsala effort and has a position with the Broad Institute.

The dream of creating a national research center in biology has been discussed for years by officials at the Karolinska Institute, Stockholm University, and the Royal Institute of Technology (KTH), all in Stockholm. But it only became a reality when the Swedish government recently offered major strategic research grants for several life sciences fields. The three institutions banded together last year to win the lion's share of the “molecular biosciences” money and then joined with Uppsala University, another grant winner, to form SciLifeLab.

By 2011, the building in Stockholm

should house more than 200 researchers. Aside from sequencing the Norway spruce, SciLifeLab intends to help Sweden take advantage of its formidable population databases and biobanks, such as the recently launched LifeGene project, which aims to gather tissue samples and track the medical histories of more than 500,000 Swedes. SciLifeLab “is what we really need to be competitive with other places and to analyze our own data,” Lindblad-Toh says.

In a leap, Sweden will create one of Europe's biggest genome centers. There are much larger centers around the world—



Center of attention. With a new facility in Stockholm (artist's illustration) that will sequence the Norway spruce genome, Sweden's Science for Life Laboratory aims to become a national life sciences center.

SciLifeLab's DNA sequencing output will be only about 40% of that of the Wellcome Trust Sanger Institute in Hinxton, U.K., for example. But Sweden is counting on an unprecedented marriage of genomics with proteomics to set it apart. Uppsala University and the Royal Institute of Technology are already home to a project, the Human Protein Atlas, that by 2014 intends to create antibodies to nearly every human protein. SciLifeLab “will be the only lab with antibodies to all human proteins and the ability to sequence human genomes. It's a unique niche,” says KTH's Joakim Lundeberg.

Scientists say they will use antibodies from the Human Protein Atlas to examine subcellular locations of human proteins, for example. And there are plans to analyze the Baltic Sea ecosystem, notes KTH's Mathias Uhlén, who heads the Human Protein Atlas and Stockholm's SciLifeLab. Lindblad-Toh says the Uppsala side will add its strengths in comparative genetics and evolutionary biology to the mix.

For now, the Norway spruce is center stage. No conifer has had its genome sequenced yet, in part because they typically are full of repetitive DNA sequences; Norway spruce and the human genome likely have comparable gene numbers, but the former is seven to 10 times bigger. “It's one of the largest, most complicated genomes to be sequenced,” says Pär Ingvarsson of Umeå University, who leads the spruce project. “We want to identify genes that control wood properties.”

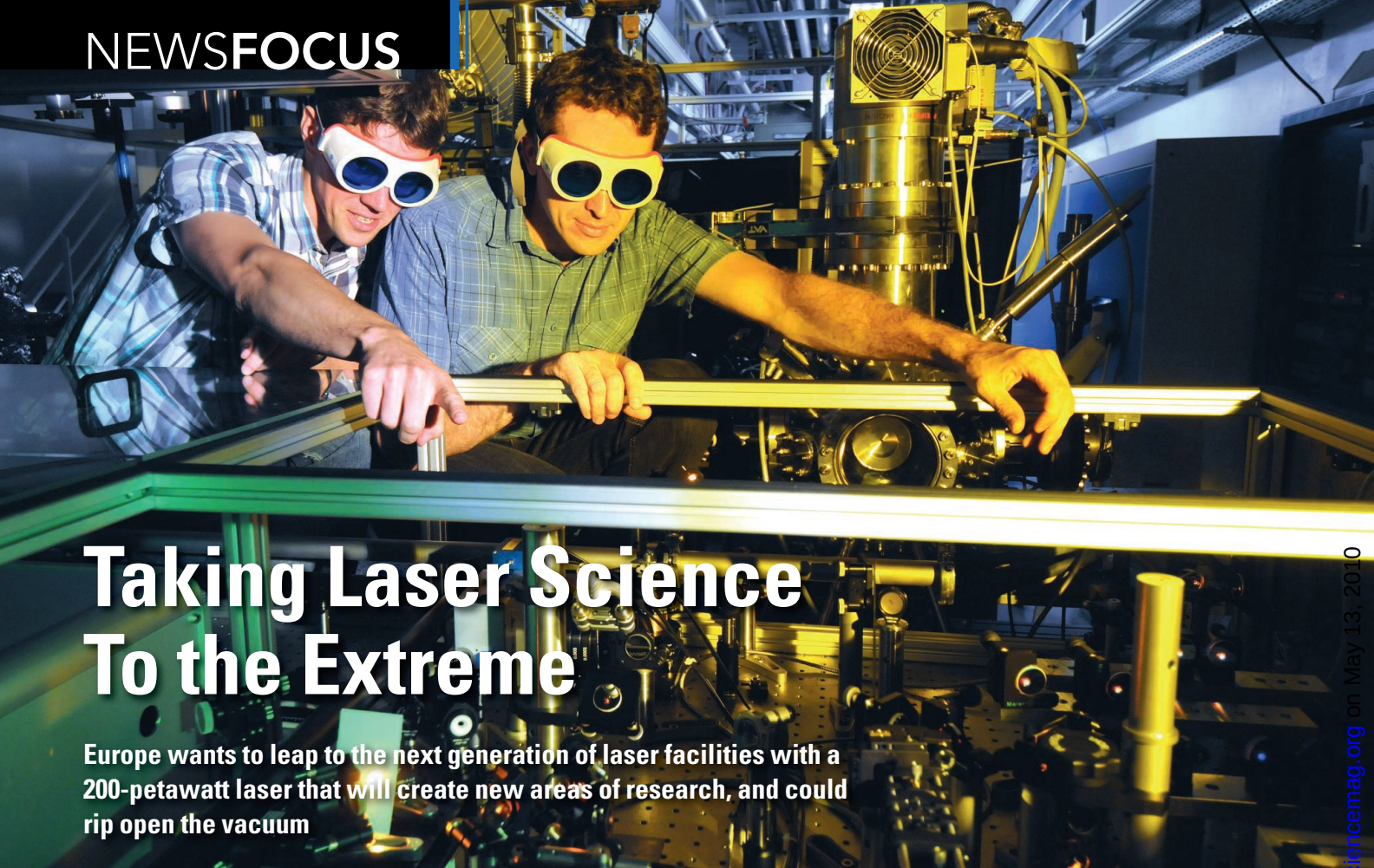
SciLifeLab's partnership with Umeå on the spruce genome is a first step toward convincing Swedish scientists outside Uppsala and Stockholm that the project is for them as well. “Most universities in Sweden are waiting to see if we can deliver useful infrastructure for them. It's up to us to prove we can,” acknowledges Uhlén.



SciLifeLab may also help the scientific communities in Uppsala and Stockholm, traditional rivals, come together. The two SciLifeLab sites intend to submit many joint grant applications, and a single person may ultimately direct the overall effort. “For the first year, we will work as two separate organizations. We're working on getting closer,” says Fredrik Sterky, site director for SciLifeLab in Stockholm. “It's a lot more powerful if we can grow as one institute with two campuses,” Lindblad-Toh says.

SciLifeLab represents a change from the past “dark decade,” when Scandinavian countries failed to use their wealth to fund strategic research efforts, notes Finland native Aarno Palotie, head of medical sequencing at the Wellcome Trust Sanger Institute. “Sweden woke up sooner than Finland or Norway.” Still, he concludes, SciLifeLab remains a gamble. The key question, asks Palotie: “Is the sum bigger than the components?”

—JOHN TRAVIS



Taking Laser Science To the Extreme

Europe wants to leap to the next generation of laser facilities with a 200-petawatt laser that will create new areas of research, and could rip open the vacuum

THE FIRST HALF-CENTURY OF THE LASER'S history has seen a constant push for higher power. Today, the Vulcan laser at the Rutherford Appleton Laboratory (RAL) near Didcot, U.K., fires pulses that have 10,000 times the power of all of Britain's electricity-generating stations added together. One of the world's most powerful lasers, Vulcan doesn't black out the entire country because its pulses are very short, less than a picosecond (10^{-12} seconds) in duration, so the energy of each pulse is a moderate 0.5 kilojoules.

Vulcan is a large machine, but over the next few years a group of European countries wants to take lasers into the realm of international big science with a facility built around a device that can produce 200-petawatt (2×10^{17} watts) pulses, 200 times the power of today's best lasers. The Extreme Light Infrastructure (ELI) isn't yet a done deal, but there is considerable political and scientific momentum behind it. That's in part because the three countries leading the project are the Czech Republic, Hungary, and Romania—all recently joined members of the European Union. "There is political pressure from the new states and the E.U. to build [research facilities] in the new states," says Wolfgang Sandner, director of the Max Born Institute in Berlin.

If ELI goes ahead, it will be the most prominent science project in Eastern Europe since the fall of the Berlin Wall. Initially split into outposts in the three countries, leading up to one mammoth laser to be built by 2017, ELI will ultimately be the Swiss army knife of laser centers. Its superfast, high-power pulses will probe the atomic nucleus and watch electrons inside atoms and molecules. By colliding pulses with various targets, researchers plan to create other sorts of radiation—electrons, protons, ions, x-rays, and gamma rays—for use in everything from cancer therapy to nuclear physics.

The ultrahigh power and intensity of pulses of ELI's final laser will produce electric fields so strong that they may alter and sense the texture of the vacuum itself, opening up new research areas for astrophysicists and particle physicists. According to quantum electrodynamics (QED), the vacuum teems with pairs of electrons and positrons that pop fleetingly into existence, briefly separate, and then recombine and disappear. The electric fields of ELI's pulses may be strong enough to pull these pairs apart before they can recombine, or at least feel the texture of this sea of virtual particles and test QED in a very direct way. We want to "drill a hole in the vacuum," says ELI proj-

ect coordinator Gérard Mourou, director of the Laboratory of Applied Optics (LAO) at Palaiseau, France.

The power of three

About 5 years ago, the E.U. called for ideas for international infrastructures to boost European research. Dozens of labs around the world already boast terawatt (10^{12} watts) lasers, and a handful, including Vulcan, can now reach petawatts (10^{15} watts). Mourou and others wanted to go even bigger. They put together a plan for a laser that would push current technology to its limits, into the hundreds of petawatts. Laser science "is ready to go to the next step, to a truly international laser infrastructure beyond the capability of a single nation," says Sandner.

Several European countries expressed interest in hosting ELI, but securing funding proved difficult until the three eastern countries realized they could apply for E.U. structural funds. These are grants given to less developed E.U. member states to build infrastructures such as roads, bridges, and hospitals, but they can equally well be spent on research facilities.

Last year, the Czech Republic, Hungary, and Romania came up with a novel plan: They would become equal partners in

CREDIT: THORSTEN NAESER/LABORATORY OF ATTOSSECOND PHYSICS AT WFO

In a flash. Researchers prepare attosecond laser experiments at the Max Planck Institute of Quantum Optics.

the project and split it in three so that each would have a facility geared to a different branch of laser science. (Institutions in another 10 E.U. nations are also involved.) The three centers would have lasers with a range of powers between 1 and a few tens of petawatts; and the decision on where to put the final 200-petawatt laser would be put off for 2 years to give researchers more time to choose the best technology.

Although laser scientists acknowledge that this makes the project more complicated, there are benefits, too. "There will be a slight increase in cost but a huge boost to local scientists [in each country]," says Sandner. If the structural funds are approved by early next year as expected, the three countries could begin pouring concrete in 2011, with a total price tag of about €750 million. "This is very, very significant. It's the first time a European infrastructure project has been built on the east side of the [former] Iron Curtain," says physicist Marius Enachescu, who is deputy secretary of state in Romania's research ministry.

The ELI facility in Hungary will focus on science using ultrashort laser pulses, just attoseconds (10^{-18} seconds) in length. Researchers began making attosecond pulses about a decade ago when they found that if they fired a femtosecond (10^{-15} seconds) laser pulse into a gas such as neon, they created higher order harmonics of the original frequency. By superposing these harmonics, they could create attosecond-scale pulses, which is just the time scale needed to discern the movement of electrons in an atom.

Researchers hope Hungary's ELI facility will enable them to carry out "pump-probe" type experiments, in which one pulse sets an atomic process in motion, then a second snaps the action a moment later like a hyperfast camera. They say they will be able for the first time to image the position in time and space of both nuclei and electrons at the subatomic scale.

The Czech branch of ELI will be a laser-based beamline facility. Many areas of science rely on beams of particles and high-energy photons from accelerators, synchrotrons, x-ray tubes, and radioactive sources. In 2000, researchers discovered that they could

also generate many of these beams by firing high-intensity laser pulses into gas jets, thin foils, and other targets. This laser strategy can produce x-rays and gamma rays, as well as pulses of electrons, protons, and ions, with a brightness and pulse length that open up new experimental possibilities. "When a laser interacts with a target, all sorts of impressive things are created. The products often cannot be produced in any other way," says John Collier, RAL's head of high-power lasers.

One possible application this ELI branch plans to explore is cancer therapy with proton or ion beams. Such beams are extremely effective for treating deep-seated tumors, but to perform such therapy a hospital now needs a particle accelerator costing tens of millions of dollars—something few can afford. Laser physicists think they can accelerate particles in a much cheaper and more compact way. When they fire a high-intensity laser pulse into a plasma, the photons' magnetic field kicks electrons in the plasma forward and these then strike a foil target. As they emerge from the other side, they drag positive ions in the pulse's wake. Such acceleration works "much faster over a shorter distance" than traditional accelerators do, says Sandner. "It's in its very early infancy, but it points a way to the next step."

The third planned ELI facility, in Romania, aims to open up a new area of laser science by probing the atomic nucleus with beams that are ultraintense—focused so that they have the maximum power per unit area. "The laser power that exists now cannot be compared with the strength of the nuclear field," says Enachescu. But he hopes that his nation's ELI outpost can change that.

It's not clear yet whether ELI's laser beams alone will be able to excite a nucleus into higher energy levels, but physicists are developing other tricks that utilize those beams to accomplish the feat.

One such scheme involves colliding a laser pulse head-on with an electron beam to produce an intense burst of gamma rays. "Then we will use that to disturb nuclei," says Mourou.

Probing the vacuum

The fourth, and at the moment least-defined, part of ELI is the final 200-petawatt laser. The uncertainty is because planners are still weighing two rival methods to stretch current technology to this new power level. All

high-power research lasers rely on amplifiers: pieces of an active lasing medium, such as glass doped with neodymium, that resemble a laser without the end mirrors. Just before a pulse is fired, the amplifier is pumped with light from another source to create a large number of excited atoms. When the pulse comes through, those atoms emit light in step with the pulse, amplifying it with extra photons. But at about a gigawatt, each pulse has so much power that it begins to damage the glass. Researchers got around this problem in the mid-1980s after Mourou and colleague Donna Strickland developed a technique called chirped pulse amplification (CPA), which reduces the peak power of a short, high-power pulse by stretching the pulse out in time, before amplifying it and compressing it again.

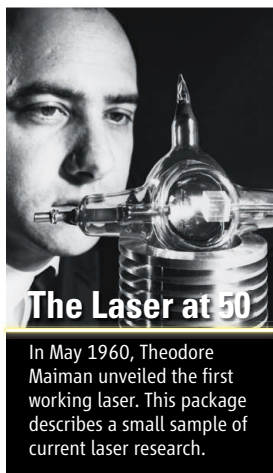
ELI may need additional strategies to increase laser power. "We're getting to the limit of CPA," says ELI Deputy Coordinator Georg Korn of the Max Planck Institute of Quantum Optics in Garching, Germany. The ELI team may consider an adaptation of CPA involving a special nonlinear crystal to transfer power from one stretched beam to another. This will be tested in a planned upgrade of Vulcan to 10 petawatts. Meanwhile, Mourou's LAO and other French labs are testing a different amplifier material, titanium-doped sapphire, by building a 10-petawatt laser.

The ELI team must eventually decide which approach to back and whether to push for even higher power or simply build 20 10-petawatt lasers and combine the beams to make one of 200 petawatts. "We have to wait for this new technology to develop. Two-hundred petawatts is so advanced that there is a need for a demonstrator," says Collier.

Even if researchers achieve that power, they will still need high-intensity beams before they can explore the vacuum. Theorists calculate that an intensity of 20^{29} watts/square centimeter (W/cm^2) will be needed to rend apart electron-positron pairs. ELI will likely be able to reach intensities of only about 10^{24} W/cm^2 , but "there are some clever ideas around, some of them not yet published," says Korn. These include using laser pulses to create intense gamma rays and probing the vacuum with them.

With enough funding, says laser scientist Donald Umstadter of the University of Nebraska, Lincoln, ELI should overcome any technical difficulties. "They've set ambitious goals to reach ideal conditions. If they do, it will be very exciting. Whenever you are going to the limits, you can expect interesting physics to emerge."

—DANIEL CLERY



The Laser at 50

In May 1960, Theodore Maiman unveiled the first working laser. This package describes a small sample of current laser research.



Dead center. The tiny device at the end of NIF's target arm is where the laser's 192 beams—and 1.8 megajoules of energy—converge.

POWER

Laser Fusion Energy Poised to Ignite

Once lasers have shown that they can spark fusion, what next?

In 1960, John Nuckolls, a physicist at Lawrence Livermore National Laboratory (LLNL) in California, was studying ways to compress and heat a capsule of hydrogen isotopes so that they would undergo fusion and generate energy. He looked at then-exotic technologies such as pulsed-power machines, particle accelerators, plasma guns, and hypervelocity pellet guns. Then in July he heard that Theodore Maiman had built the first successful optical laser at Hughes Research Laboratory in Malibu, California. Within days Nuckolls was working out how to implode his fuel capsule with a laser.

Today, the direct result of Nuckolls's initiative is sitting in a cavernous hall 10 stories high and large enough to cover three football fields: the \$3.5 billion National Ignition Facility. NIF was completed in 2009, and test firing of its giant laser has been going on since last summer. Shots aimed at igniting fusion fuel could begin later this year. Nuckolls—now a director emeritus of LLNL—and the rest of the fusion community are waiting anxiously to see if NIF's laser, which produces the world's highest energy pulses, really can ignite a tiny version of the sun. "NIF will prove that ignition takes place at laboratory temperatures, irrespective of the heating method," says Hiroshi Azechi, director of the Institute of Laser Engineering at Osaka University in Japan. Laser, or inertial, fusion energy "needs ignition to be successful," says

Mike Dunne, director of the Central Laser Facility at the Rutherford Appleton Laboratory in Didcot, U.K. "There's a lot of belief. The question is when?"

Ignition alone, however, is not enough to prove that inertial fusion is a viable energy source. For all of its importance as a demonstration project, NIF would make a terrible power station. To give NIF the best chance of success, LLNL researchers built a laser designed to produce one-off pulses with the highest energy currently possible: 1.8 megajoules in a flash lasting about 20 nanoseconds. The tradeoff is that NIF can fire only two or three times per day. If everything goes according to plan and each fusion burn produces 100 megajoules of energy, three shots a day would produce about 3.5 kilowatts of heat—enough to power one typical American home. Researchers estimate that an inertial fusion power station would need to explode a fuel capsule about 10 times a second, a feat no high-energy lasers can now achieve.

Other obstacles to inertial fusion power include manufacturing the tiny fuel capsules filled with hydrogen at a rate of about a million a day; feeding them into the path of the laser beams 10 times a second; and building a vessel to contain the explosions and absorb the neutrons produced in fusion while converting their energy into heat to generate electricity. Even as they look for-

ward to news from NIF, researchers at projects in Japan, Europe, and the United States are already exploring steps toward practical laser-driven fusion power, should the laser in Livermore succeed. Which projects offer the best hope, and whether all can survive tight funding, remains unclear.

Pulse power

Key to any attempt at inertial fusion energy is building the laser. Industry has had decades of experience with high-power lasers of about 100 kilowatts, used for cutting, for example. Fusion could manage with 50 kilowatts on average, Dunne says, but it needs to be delivered not as a continuous beam but in very short pulses each with high energy. "The challenge is to marry short pulses with high average power," says Dunne, a combination he thinks will take about 5 years to achieve.

High-energy lasers such as NIF's start with a relatively low-power beam and pass it through a series of amplifiers that steadily boost its power. An amplifier is a large piece of lasing material pumped with light from another source to create a population of excited atoms. When the beam passes through, the atoms emit photons in step with it, increasing its power. NIF's amplifiers are made from glass doped with neodymium; the pumping is done by thousands of xenon flash lamps powered by a bank of capacitors that can store 400 megajoules of energy. That's fine for NIF but not for a true laser-fusion power plant: Flash lamps must be cooled between shots and their capacitors recharged, and they are less than 1% efficient from wall plug to light. A power plant would need something at least 10% efficient—which, most agree, means solid-state laser diodes.

Laser diodes, developed for the telecommunications industry to send light down optical fibers, are "getting better all the time," says NIF Director Ed Moses. A diode-pumped laser would have arrays of diodes wrapped around the amplifier material to provide intense flashes of light at a high repetition rate. NIF researchers have built a demonstrator laser called Mercury that Moses says can produce the required repetition rate but has a narrow aperture and hence low power. Such lasers "have come a huge distance over the last decade," he says.

Livermore's proposal for a post-NIF fusion energy project, called LIFE (Laser Inertial Fusion Engine), would produce up to 500 megawatts of power. LIFE would be "part of a continuum" from NIF, Moses says. The whole laser system would "have the look

and feel of NIF but operate at 10 hertz rather than hours.” One potential stumbling block is the high cost of laser diodes. “We could operate it now. We can show the performance but not the economics,” Moses says.

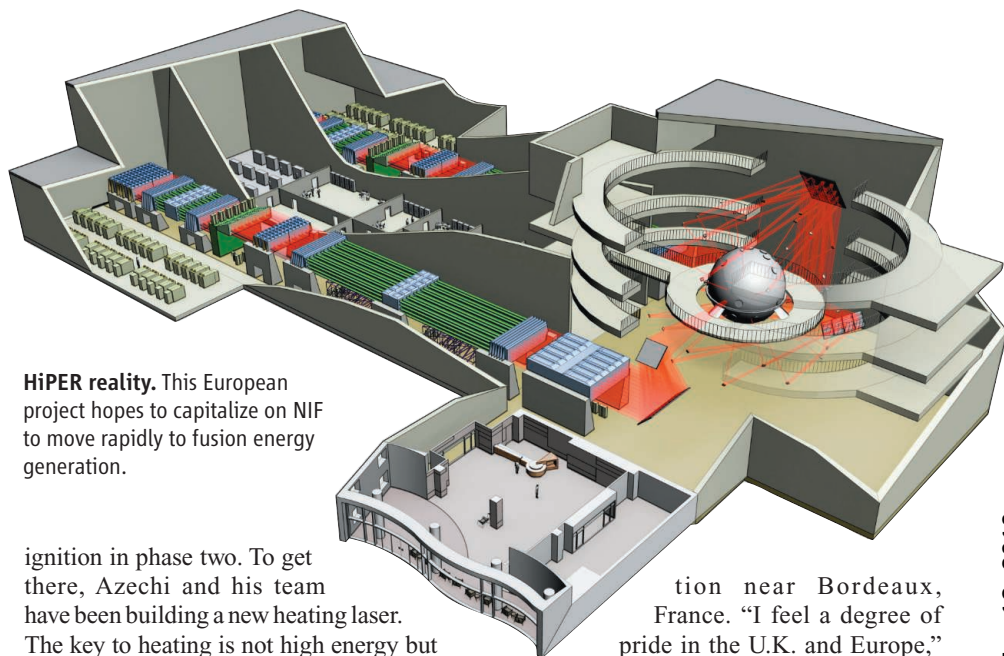
Others, however, say leaping straight from NIF to a power plant is the wrong way to go. NIF’s and LIFE’s lasers, they point out, have to be very high energy because they are doing two jobs: compressing the fuel capsule and heating it to spark ignition. Some researchers predict that if you separate those functions and build two lasers, each optimized for a different job, you would have a much more efficient system. The compressing driver laser would be about one-tenth of the energy of NIF’s, and the heating laser less energetic still.

This approach, known as fast ignition fusion, is under investigation at several centers, most aggressively at Osaka University in Japan. About 5 years ago, a team there led by Azechi proved the principle of fast ignition by first compressing a hydrogen-filled capsule to 600 times the density of a normal liquid using Osaka’s GEKKO-XII laser and then raising its temperature to about 10 million kelvin with a second laser. The experiment produced fusion reactions but not ignition. On the strength of that result, the team embarked on the Fast Ignition Realization Experiment (FIREX).

The first phase of FIREX aims to reach the ignition temperature of more than 60 million kelvin. Then the team will attempt



One-two punch. The FIREX project laser will provide the heating pulse to spark fusion.



HiPER reality. This European project hopes to capitalize on NIF to move rapidly to fusion energy generation.

ignition in phase two. To get there, Azechi and his team have been building a new heating laser.

The key to heating is not high energy but high power in very short pulses. The FIREX heating laser will be able to produce pulses with petawatt (10^{15} watts) power lasting just 10 picoseconds (10^{-11} seconds). But the process has not been easy. “The power is so high, everything breaks,” says Azechi.

To keep the high-power pulses from destroying amplifiers, laser researchers use a technique called chirped pulse amplification: They take a short pulse, stretch it out in time using a diffraction grating, which reduces its peak power, pass it through an amplifier, and then use another grating to compress it back into a short pulse, now with higher power. “We need huge gratings with very high precision,” says Azechi. “It’s a technical challenge that has taken almost 5 years.” The team has one of the laser’s four beamlines working and will complete the rest in the next 1 or 2 years, Azechi says. But he adds that he’s worried about funding for fusion in Japan. “With the technology, I’m very optimistic. Politically, it’s difficult. There are so many projects in this field.”

In Europe, fusion researchers are betting on fast ignition to jump straight to a demonstration power-producing reactor called the High Power laser Energy Research (HiPER) facility. The project—a collaboration involving labs in 11 countries led by Dunne’s Central Laser Facility—has already completed a design and is in the middle of a 3-year preparatory phase during which a consortium, funders, and industry involvement will all be put in place. The plan is to begin construction in mid-decade if all goes well at NIF.

Work to develop HiPER’s lasers is going on all over Europe, and a test-bed fast ignition laser called PETAL is under construc-

tion near Bordeaux, France. “I feel a degree of pride in the U.K. and Europe,” says Dunne, who currently leads the project. “We’re in a strong position with the technology and strategy. We’re at the leading edge of this field.” Moses, however, thinks fast ignition isn’t as close to viability as some believe. “LIFE uses low-risk technology. Fast ignition and other schemes are more futuristic, but they should be supported,” he says.

All for one?

In the United States, meanwhile, the recent focus on NIF has largely diverted the government’s attention from what comes next. (LIFE is an in-house LLNL project.) That is beginning to change. Moses says both Energy Secretary Steven Chu and one of his deputies, Steven Koonin, have visited the lab in recent months to learn about NIF and LIFE. Koonin has tasked the National Academy of Sciences to report on the next steps to achieve inertial fusion energy. “Our goal is to create a U.S. national effort, perhaps with international participation,” Moses says.

If success at NIF sets the field moving again, the question of cooperation is bound to arise: Does the world need three projects doing roughly the same thing? “We’re interested in collaborations. How it develops will be very interesting,” says Moses. Dunne, who will soon leave the United Kingdom to become director of the LIFE project at Livermore, says he will work to merge the LIFE and HiPER teams. “It’s the best way to achieve a result in the shortest possible time,” he says. “It’s not about whether laser fusion will happen, but who will make it happen.”

—DANIEL CLERY

NANOLASERS

Ever-Smaller Lasers Pave the Way For Data Highways Made of Light

New materials and techniques are bringing researchers close to a once-unthinkable goal: optical devices tiny enough to work hand in hand with electronic circuits

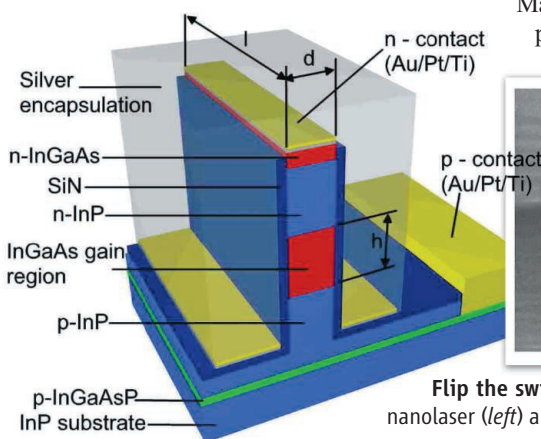
The dream of optical computing—replacing electronic devices with much faster ones based on light—has tantalized scientists for generations. Nowadays, computer circuitry has grown too complex to be replaced wholesale. Instead, researchers talk about using lasers and other optical components as high-speed data highways between specialized electronic processors on chips. So far, lasers and other optical components have been far too big to make this integration possible. “It’s hard to integrate the two technologies when the optical devices are 1000 times larger than the electronic devices,” says Cun-Zheng Ning, a physicist at Arizona State University, Tempe.

But the gap is narrowing. In recent years, researchers around the world have married traditional optical materials with metals to create lasers a mere tens of nanometers thick. Just last month, two groups reported making lasers ultrasmall in all three dimensions. “That was almost unthinkable before,” says Peidong Yang, a chemist at the University of California (UC), Berkeley. “There has been a lot of progress.”

Shrinking conventional lasers long seemed all but impossible. Traditional devices produce laser light by sending photons through an optical cavity made from a “gain” material and bouncing them back and forth with tiny mirrors. Along the way, the photons cause energized electrons in the gain material to release their energy as additional photons of light with the same wavelength. Those released photons then stimulate the release of still more photons. As the avalanche grows, some of the light is allowed to leak out of one of the mirrors to produce a tight beam of photons whose waves all travel in lockstep.

In most lasers, the gain region of the optical cavity is micrometers to meters in length. Photons of light can’t be confined to spaces smaller than half their wavelength; otherwise they leak out. Making a smaller device lase is like pouring water into a sieve. “A classical [laser] cavity does not work,” Yang says. As a result, conventional lasers and other optical components for

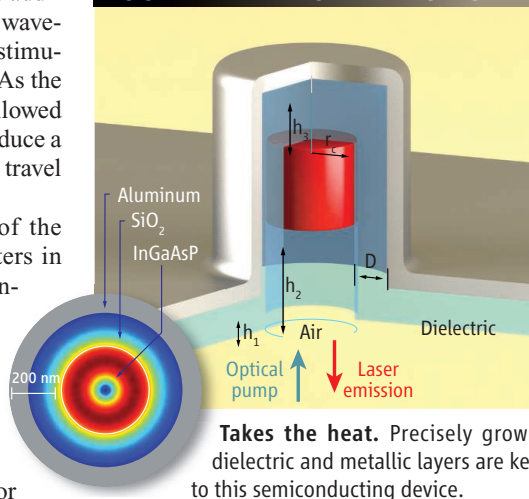
visible light can’t be much smaller than 200 to 300 nanometers across, and most conventional optical components are much larger than that. Current transistors and other electronic devices, by contrast, include features that measure just tens of nanometers across.



Flip the switch. Structure of an electrically activated nanolaser (left) and SEM image of its semiconducting core.

The effort to shrink lasers took an important step in 2001, when Yang and his colleagues reported making the first laser from a nanowire, a whisker of zinc oxide as thin as 100 nanometers across (*Science*, 8 June 2001, p. 1897). Still, even though these lasers were ultrasmall in two dimensions, they were 4 to 10 micrometers long, so that photons traversing the optical cavity would have room to build up enough gain to lase.

ROOM TEMPERATURE NANOLASER



Takes the heat. Precisely grown dielectric and metallic layers are key to this semiconducting device.

Glimmers of a breakthrough came in the late 1980s and 1990s from researchers studying the way light interacts with electrons in metals. Metals are strong light absorbers. But researchers found that shining a light beam at the interface between a metal and a non-conductive (or “dielectric”) medium such as glass caused mobile electrons at the interface to oscillate back and forth at the same frequency as the light but with a much shorter wavelength. That discovery offered the hope that by coupling a laser’s gain medium with a dielectric-metal interface, they could effectively squeeze the wavelength of the light produced by the device and thus make the overall device smaller.

Making such lasers took several years to pull off, but recent progress has been

brisk. In 2007, researchers led by Martin Hill, an electrical engineer at Eindhoven University of Technology in the Netherlands, were the first to show that a semiconductor surrounded by a thin metal structure could lase. More recently, Hill, Ning, and colleagues reported in the 18 June 2009 issue of *Optics Express* that they had made more complex lasers from pillars of indium-gallium-arsenide semiconductor as thin as 90 nanometers surrounded by a 20-nanometer-thick silicon nitride dielectric and a silver casing. Hill and his colleagues showed that the ultrasmall structure would lase when fed electricity—an impressive feat, as most new lasers need to be “pumped” with photons from another laser. Shaya Fainman, an electrical engineer at UC San Diego, calls the new work “pioneering” and says he’s “a little jealous” of the paper. But he notes that the device works only at cryogenic temperatures. “To make it practical, it has to be electrically pumped and work at room temperature,” Fainman says.

Last month, Fainman’s UCSD group reported that they had achieved this higher temperature operation. In their device, the UCSD researchers created an indium-gallium-arsenide-phosphide cylinder for

Smallest of the Small?

Lasers have been shrinking ever since they were invented 50 years ago. But the 44-nanometer "spaser" will be hard to beat.

The device, reported online in *Nature* on 16 August 2009 by researchers at Norfolk State University in Virginia; Purdue University in West Lafayette, Indiana; and Cornell University, is akin to a conventional laser. In traditional lasers, light bounces between two mirrors, stimulating a “gain” medium to release more photons of light at the same wavelength to produce a tight, concentrated beam. In their surface-plasmon laser, or spaser, Mikhail Noginov, a physicist at Norfolk State, and his colleagues shine a conventional bluish-green laser beam into a suspension of nanoparticles, each with a gold core surrounded by a layer of sodium silicate and an outer silica shell containing dye molecules. When the gold is excited by the laser photons, its surface starts to dance with collective oscillations of electrons, known as surface plasmons. The plasmons tickle the dye molecules to release more plasmons at the same wavelength, causing the device to emit green laser light.

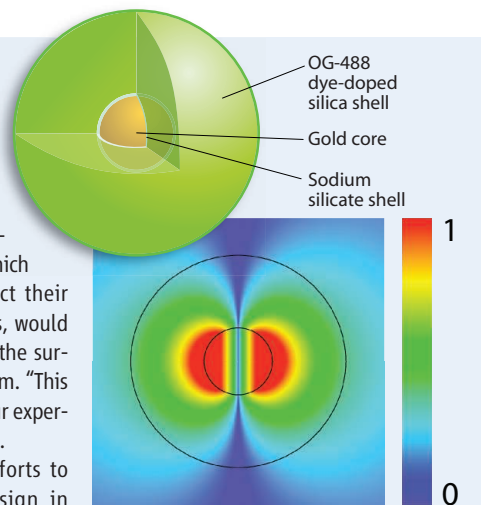
"It's impressive work," says Cun-Zheng Ning, a nanolaser expert at Arizona State University, Tempe. But other outsiders are more cautious. "There are questions," says one physicist, who asked not to be identified. Among the issues is a concern that the lasing could be a collective response from

clusters of nanoparticles rather than from individual particles. Noginov and his colleagues respond that aggregation of particles, which can potentially affect their stimulated emissions, would also cause a shift of the surface plasmon spectrum. "This was not the case of our experiment," Noginov says.

Noginov says efforts to use this spaser design in high-speed computing could confront a greater challenge. The organic dye molecules used in the nanoparticles are fragile. To make a device robust enough for use in computing technology, Noginov says, researchers will likely need to make new all-inorganic-based spaser nanoparticles, a goal the field now faces.

—R.F.S.

-R.F.S.



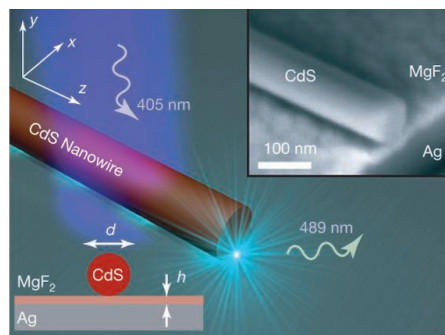
Pinpoint. Laser-zapped nanoparticles (*top*) generate surface plasmons, which multiply and reemit laser light at a different color.

the gain material. They then enveloped all but the bottom of the cylinder with a thin dielectric layer of silica and an outer layer of aluminum, leaving the bottom exposed to the air. During operation, the researchers fire a 1.064-micrometer laser beam at the open part of the pillar. The device uses that energy to create laser light at a longer 1.43 micrometers, which is then reemitted out of the uncoated bottom region of the pillar. Fainman says that they got the device to operate at room temperature by precisely engineering the width of the dielectric and metal layers. Now they must get the device to work when fed with electricity instead of a separate laser beam. "We are working on it," Fainman says.

A very different-looking device was recently reported by physicist Xiang Zhang and colleagues at UC Berkeley. In the 1 October 2009 issue of *Nature*, Zhang's group reported creating a plasmonic laser using a long, thin cadmium sulfide (CdS) nanowire placed atop a 5-nanometer-thick dielectric layer of magnesium fluoride, which in turn rests on a silver surface. The device starts out much like Yang's original nanowire laser: First, the 10-micrometer-long CdS nanowire is pumped with a blast from a separate laser. Those photons are absorbed in the semiconductor, creating energetic electrons paired with electron vacancies called holes. When those electron-hole pairs recombine, they emit photons at a longer wavelength.

Some of those reemitted photons, Zhang explains, bounce back and forth in the nano-

wire as in a traditional nanowire laser. Most, however, strike the metal-dielectric interface and generate oscillating electrons known as plasmons, which can travel like ripples in a pond. Typically, such plasmons would travel only a very short distance in a metal before dissipating. But by sandwiching the magne-



By a whisker. Electron waves called surface plasmons surf beneath this nanowire, stimulating more plasmons and eventually laser light.

sium fluoride dielectric between the semiconducting CdS nanowire and the metallic silver, Zhang says, the researchers can channel the plasmons into it. Because the dielectric material doesn't absorb plasmons, they can travel long distances back and forth. Tails of the plasmons' electromagnetic waves extend into the nanowire and prod it to emit more photons at the same plasmonic frequency, creating the hallmark avalanche effect of a laser. One end of the nanowire is slightly more leaky, and plasmons emerging from the structure produce a beam of greenish blue light.

In a paper posted on the arXiv preprint server* on 23 April, Zhang and colleagues report creating a related device that improves on the one they described in *Nature*. The new device works at room temperature instead of cryogenic temperatures, and it shrinks the longest dimension of the laser down to a single micrometer. Zhang says his team is now working on an electrically pumped version of their plasmonic laser.

Progress is coming from other angles as well. Last year, for example, researchers in Virginia, New York, and Indiana reported making a laser from 44-nanometer-sized particles (see sidebar, above). And last month, Yang's group at UC Berkeley reported having made an ultraviolet laser smaller than the wavelength of light in all three dimensions using a different technique known as "whispering gallery mode" lasing. Michael Haney, a physicist who manages research in this area for the Defense Advanced Research Projects Agency in Arlington, Virginia, says such progress shows that the goal of making lasers on the scale of electronic devices is "advancing very quickly." The next challenges, Haney says, are to get usable light out, make the lasers electrically activated, and assemble them in large arrays that can be integrated with electrical components. Those hurdles won't be easy to clear. But the progress achieved thus far makes the prospects much brighter.

—ROBERT F. SERVICE

*<http://arxiv.org/abs/1004.4227>

OPTICS

Putting Light's Light Touch to Work As Optics Meets Mechanics

Forces exerted by light can set tiny objects aquiver, a phenomenon scientists hope to harness in the burgeoning field of cavity optomechanics

The effect appeared unexpectedly. In 2005, Kerry Vahala, an applied physicist at the California Institute of Technology (Caltech) in Pasadena, and his team were experimenting with microtoroids—little glass disks that resemble dinner plates with fat, rounded rims. Light can circulate around the rim, and each microtoroid resonates with laser light of a specific wavelength and frequency, just as an organ pipe rings at a distinct pitch. But the researchers found that the light passing through such a doohickey also warbled up and down in intensity at a much lower frequency (*Science*, 15 July 2005, p. 366).

Eventually, the researchers found the source of the radio-frequency warbling: Pressure from the light within the toroid was making it vibrate. “We were not looking for this [mechanical effect] at all,” Vahala says. “There was no precedent for it aside from some theoretical speculation.”

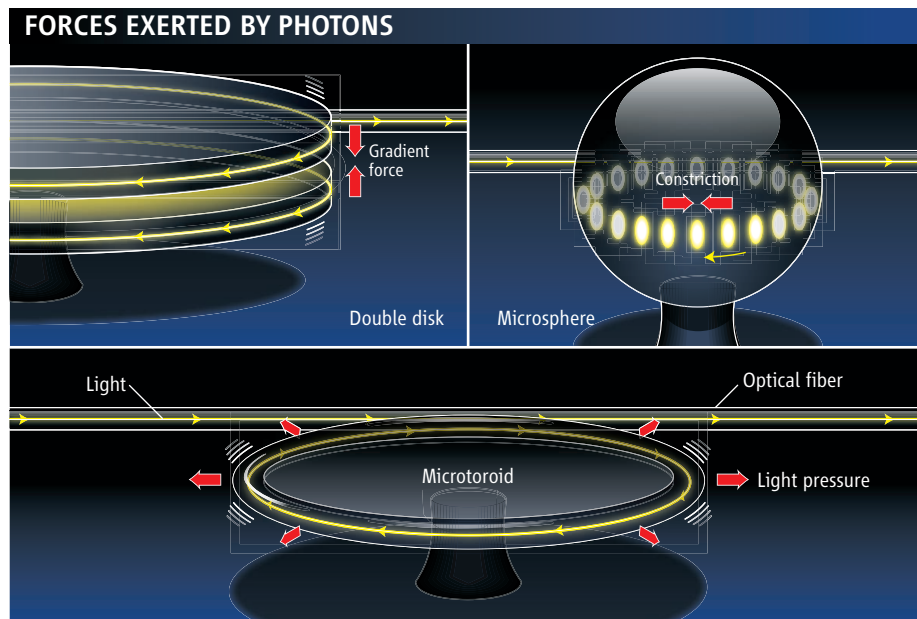
That light-and-motion connection has spawned a new field. Scientists have long used laser light to move atoms and molecules—Nobel laureate and U.S. Secretary of Energy Steven Chu developed “optical tweezers” that can uncoil DNA. But researchers are now using light to control the motion of larger humanmade objects built of materials like glass or silicon. A device must be both a fine “optical cavity” that rings with light and an outstanding “mechanical resonator” that, like a pitchfork, vibrates readily at a precise frequency. Researchers can then use light to control vibrations or vice versa.

Such “cavity optomechanics” might be used to coax vibrating machines to hum to the odd rules of quantum mechanics, which ordinarily govern the realm of atoms and molecules. More practically, optomechanical widgets might process optical signals or serve as frequency standards. One team has even used them to make a kind of laser for vibrations. Other exciting results are sure to come, says Oskar Painter, an applied physicist at Caltech: “This wave hasn’t crested yet.”

All aquiver

Light can shake an object in several ways. In Vahala’s gizmo, light’s pressure does the trick.

A toroid will resonate with light bled in from an optical fiber if its circumference equals a multiple of the light’s wavelength. As light accumulates, its pressure stretches the toroid (see figure, below). But that stretch spoils the resonance, so the circulating light wanes, the pressure eases, and the toroid shrinks. The cycle of expansion and contraction repeats at a rate set by the stiffness of the disk, typically millions of cycles per second, and makes the light leaving the cavity warble.



Give it a shake. Light can set a micrometer-sized object vibrating by exerting pressure (bottom), creating sideways pushes or pulls (top left), or causing material to constrict.

Light exerts a sideways “gradient force” as an insulating material feels a pull to where the light is most intense, like an actor who craves the spotlight. For example, if one toroid lies atop another, then gradients in the light leaking between them can pull their edges together or push them apart, as Michal Lipson, an applied physicist at Cornell University, and colleagues reported online 15 November 2009 in *Nature*. If the toroids are springy, then that force can also trigger their edges to flap up and down in a cycle akin to the one pressure causes in a single toroid, as Painter reported on 31 August 2009 in *Physical Review Letters*.

Light can also trigger vibration by making a material contract. Tal Carmon, an applied physicist at the University of Michigan, Ann Arbor, and a colleague circulated light in a 100-micrometer-wide glass sphere. In a chicken-or-egg process called stimulated Brillouin scattering, the light set off a vibration moving in the same direction, while the vibration “scattered” light into a second wave of a longer wavelength going the other way. The overlapping light waves created a moving pattern of bright spots that caused the glass to contract and amplified the vibration, as the team reported on 19 March 2009 in *Physical Review Letters*.

Light is a wimpy mover and shaker; a laser beam with a power of 1 watt exerts a force of just 6 nanonewtons. To respond to such feeble shoving, an object must resonate with light, so that each photon circulates in it thousands of times, greatly enhancing the

photon’s mechanical effect. The device must also ring with vibrations, so a small force repeatedly applied can cause a large movement. Only recently has it been possible to combine those properties. “These structures weren’t around 15 years ago,” Vahala says.

Cool it!

Now that the gadgets exist, physicists want to see what they can do. Many are trying to coax them to wiggle quantum mechanically. According to quantum theory, a vibrating object, or “oscillator,” can absorb energy only in discrete “quanta” whose size is set by the vibration frequency. The object can’t

stand still, either; it will always possess an inextricable half-quantum of energy and hum with “zero-point motion.” Physicists hope to spot that unquenchable quaking as a step toward more-bizarre states of motion, such as one in which an oscillator is in two places at once.

Each quantum is minuscule. So to remove every possible one and reach the “ground state,” physicists must cool an oscillator to nearly absolute zero. Ironically, Vahala’s observation that light can trigger vibrations opens the way to cooling an oscillator by shining a laser on it. The trick is to reverse the process, says Tobias Kippenberg of the Swiss Federal Institute of Technology, Lausanne. “The fact that you could use this for cooling was immediately clear,” he says.

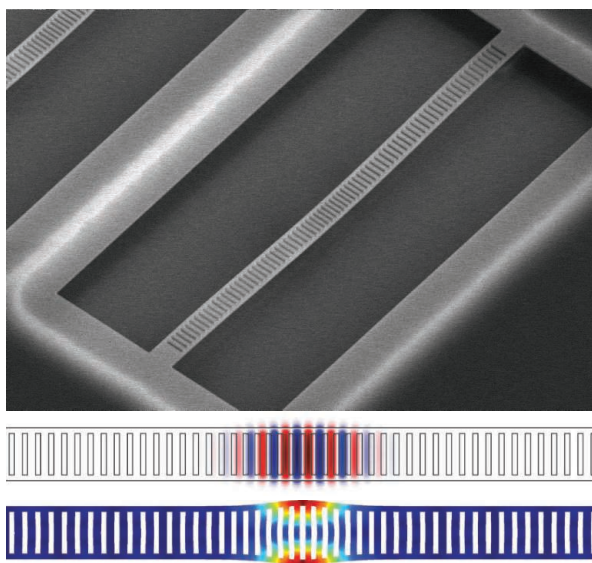
To trigger a vibration in, say, a toroid, the light’s frequency must be tuned slightly higher than the toroid’s resonant frequency. To still a vibration, the light’s energy and frequency must be lowered—optimally, by the frequency of the vibration. To enter the toroid, a photon must then make up the energy it lacks by absorbing a quantum of energy from the toroid. The scheme works because, thermodynamically, “a laser beam is a very cold object,” says Jack Harris of Yale University. “A good laser is at 0 kelvin.”

Last year, three groups used laser cooling to reduce an oscillator’s energy to a few dozen quanta. However, they were beaten to the ground state by a team that simply chilled an oscillator with a very high frequency—and hence large quanta—in a liquid-helium refrigerator (*Science*, 29 January, p. 516). But optical methods still have advantages, says Markus Aspelmeyer of the University of Vienna. “That’s why the result has only encouraged us,” he says.

Laser cooling alone might chill an oscillator to the ground state, Aspelmeyer says. That might make it possible to run an array of oscillators not deep within a helium refrigerator but on a more-accessible optical table to process information encoded in the quantum state of light. And optical methods might do things that other methods can’t, Harris says. For example, he has devised an optical scheme to watch the quantum jumps when an oscillator absorbs individual quanta. Optomechanical devices could also generate exotic quantum states of light, Harris says, such as “squeezed” states that suppress the random lumpiness of the photons in the light.

Other vibes

Optomechanical gizmos might serve myriad other purposes, too. “I’m very excited about the applications of these structures,” says Cornell’s Lipson. A double-toroid makes a low-power optical switch, she says, as one laser beam can change the size and optical frequency of the thing and control the passage of another beam. A jiggling toroid makes a frequency standard that’s immune to electrical interference. It works as a “down-



Conceptual bridge. The array of holes in a bridgelike optomechanical crystal (top) traps light (middle) and vibrations, transferring energy from one to the other.

converter” to extract a radio signal encoded in an optical carrier beam, as Mani Hossein-Zadeh, an applied physicist at the University of New Mexico, Albuquerque, reported on 15 February 2008 in *IEEE Photonics Technology Letters*.

Meanwhile, other researchers are working on more speculative ideas. In the 1990s, physicists pioneered “photonic crystals,” materials riddled with holes through which photons travel like the electrons in a crystalline solid, with only certain energies allowed. Similar holey “phononic crystals” control vibrations in the same way. Now, Caltech’s Painter and his colleagues have combined the two, as they reported online 18 October 2009 in *Nature*.

Their optomechanical crystal consists of a ladderlike beam of silicon about 30 micrometers long and 1.3 micrometers wide (see figure, above). The spacing between rungs shrinks slightly near the beam’s middle, creating a “defect” that traps both light of a certain frequency and vibrations shaking at about 2 billion cycles per second. Again, light forces set off vibrations, which reveal themselves in a warbling of the light.

Next, the team used optical gradient forces to couple the jiggling of two parallel bridges, tuning their combined motion to a so-called dark state that neither emits nor absorbs light, as they reported 7 February in *Nature Photonics*. That feat raises the possibility of snaring a light pulse in that dark state. “We’ve got exactly what we want, which is the ability to store a photon pulse and then release it,” Painter says. Physicists had previously stopped light in atomic vapors, but optomechanical crystals are easier to control and could be used for all-optical quantum information processing, he says.

Physicists have even fashioned the equivalent of a laser for vibrations. In a laser, a light-emitting material sits between two mirrors. The atoms in it are boosted from their ground state to a more energetic, excited state by, say, applying a voltage, and as they “de-excite,” the atoms emit photons, which bounce between the mirrors. Passing through the material, the photons trigger excited atoms to emit more photons of the same wavelength and direction. Such “stimulated emission” produces a torrent of identical photons that is the laser beam.

Twisting that idea, Caltech’s Vahala amps up vibrations using jumps between optical states of two toroids placed side by side. As light bleeds from one to the other, their resonances meld into a lower frequency one and a higher frequency one, just as the electronic orbitals of two identical atoms meld and split when the atoms form a molecule. Adjusting the toroids’ spacing, the researchers tune the frequency difference to equal the frequency at which one toroid shakes, as they reported 22 February in *Physical Review Letters*. When they then pump in enough high-frequency light, the vibrations amplify themselves by stimulating jumps to the low-frequency optical state.

Where all these efforts will lead remains to be seen. Some researchers say that, technologically, optomechanical devices have yet to do anything that electronics can’t do. “I don’t see anything that’s a breakthrough at this point,” Hossein-Zadeh says. However, the tiny machines may find uses in basic research in various fields of physics. “These little widgets will be a part of science, even outside of what we call cavity optomechanics,” Vahala says. Now that physicists can shake things up with light, likely they’ll find ways to use that trick.

—ADRIAN CHO



LETTERS

edited by Jennifer Sills

China's Environmental Civilian Activism

IN THE POLICY FORUM "CHINA'S ROAD TO SUSTAINABILITY" (2 APRIL, P. 50), J. LIU OVERLOOKS AN important cultural force. China's worsening environmental conditions have catalyzed a spirit of environmental civilian activism.

For example, in 2003, a consortium proposed erecting 13 dams on the Nujiang River. China's environmental nongovernmental organizations and scholars launched a protest campaign through the Internet and newspapers. The critics argued that as reservoirs behind the dams filled up, flooding and landslides would imperil habitats. In response, Premier Wen Jiabao suspended the dam project pending an environmental review in 2004 (1).

The landmark of environmental civilian activism occurred in Xiamen City in 2007. The local government supported construction of a \$1.4-billion paraxylene plant near the center of the city. Information about the environmental impact of this project was not made available to the local residents. The people of Xiamen City were outraged when—through cell phone messages and the Internet—they learned of the plant's environmental risks. A phone text message was circulated among Xiamen citizens in late May calling for a "collective walk" (demonstration). On 1 June 2007, more than 1000 citizens gathered in front of the municipal building to protest. The demonstration forced the local government to cancel the largest industrial project in the history of Xiamen (2).



Protests. Chinese citizens protest against the planned Guangzhou trash incinerator in 2009. Their banners read, "Oppose the trash incinerator."

by the authorities. (In earlier years, standard practice was to obey Beijing-based experts in environmental protection.) In addition, the self-organized middle class forced the local government to open discussion by Internet. By seizing the opportunity for an open discussion, the newly empowered locals took to the streets to protect their environmental rights (4).

Recent years have witnessed an impressive growth in environmental protests in China. The number of petitions and mass public protests related to environmental issues has increased by 30% per year in the past few years, although the number of petitions lodged with the Chinese government has dropped (5).

The current environmental civilian activism movements have several common characteristics: (i) They are confined to one specific geographical space. (ii) Their goal is protecting the environment, rather than political rights or commercial interests. (iii) They focus on a specific pollutant, rather than general environmental degradation. The local nature of the movement enables the organization of a large number of citizens with little effort in a very short time. Given more open social and political conditions and the increasing size of the middle class in

The burgeoning middle class has become the driver of environmental civilian activism. For example, operation of the Likeng trash incinerator in Guangzhou City started in 2005 without any protest, although local farmers worried about health risk (3). In contrast, the proposed Panyu trash incinerator in Guangzhou City in 2009 triggered protests that were led by the middle class (4), who used science-based evidence to openly challenge prevailing notions formulated

China, environmental civilian activism will certainly be a key driver in China's transition to sustainability.

QIANG WANG

Qingdao Institute of Bioenergy and Bioprocess Technology, Chinese Academy of Science, Qingdao, 266101, China; Guangzhou Institute of Geochemistry, Chinese Academy of Science, Guangzhou, 510640, China. E-mail: tswangqiang@yahoo.com.cn

References and Notes

1. R. Stone, *Science* **319**, 1604 (2008).
2. F. Zeng, Z. Jiang, "Annual figures: Xiamen citizen in PX protest," *Southern People Weekly Magazine* (18 January 2008); www.infzm.com/content/4640 (in Chinese).
3. J. Chen, "Some farmers near Likeng trash incinerator died of cancer," *China News Weekly* (4 December 2009); reproduced at http://news.sina.com.cn/c/sd/2009-11-26/111519132701_2.shtml (in Chinese).
4. M. Moore, "China's middle-class rise up in environmental protest," *Telegraph* (23 November 2009); www.telegraph.co.uk/news/worldnews/asia/china/6636631/Chinas-middle-class-rise-up-in-environmental-protest.html.
5. S. Zhou, "Creating environmental supervision and environmental law enforcement to cope with new task and new challenges" (Ministry of Environmental Protection of the People's Republic of China, 2008); www.mep.gov.cn/info/ldjh/200804/t20080425_121679.htm (in Chinese).
6. The views expressed in this paper are the author's own and not necessarily those of QIBET-CAS, GIG-CAS.

I thank B. Jong for comments and linguistic support.

Effects of China's Economic Growth

IN A RECENT POLICY FORUM, J. LIU REVIEWED "China's road to sustainability" (2 April, p. 50). Liu focused on population growth and an increase in the number of households, but he failed to adequately address the most important socioeconomic driver behind environmental degradation in China: rapid economic growth that is not offset by efficiency improvements (1, 2).

In China, exports and capital investments contribute significantly more to gross domestic product (GDP) than household and government consumption combined (3), and this also holds true for emissions (1, 2). From 2002 to 2005, the production of exports was responsible for 50% of the growth in carbon dioxide emissions and capital formation was responsible for 35%;



Global warming
and lizards

832



Mobile small
RNA

834

household and government consumption contributed 15% (4, 5).

Liu discusses possible population control and household size, but a more dominant issue in terms of population dynamics is the migration from rural to urban areas (6). From 1990 to 2007, the urban population increased by 292 million, whereas the rural population decreased by 116 million (3). Urban dwellers, even if migrants from rural areas, have a higher income (3) and hence higher energy use and environmental impacts (2, 6).

A key challenge for China is to continue strong economic growth while minimizing environmental impacts. Reductions in emis-

sions per unit of GDP are unlikely to reduce total emissions if economic growth continues (1). China will need to combine aggressive domestic policies with international support to reverse the current growth in coal-dominated energy use and emissions.

GLEN P. PETERS,^{1*} DABO GUAN,²
KLAUS HUBACEK,^{3,4} JAN C. MINX,^{5,6}
CHRISTOPHER L. WEBER⁷

¹Center for International Climate and Environmental Research—Oslo (CICERO), Norway. ²Cambridge Centre for Climate Change Mitigation Research, Department of Land Economy, University of Cambridge, UK. ³Sustainability Research Institute, School of Earth and Environment, University of Leeds, UK. ⁴Department of Geography, University of Maryland, College Park, MD 20742, USA. ⁵Stockholm Environment Institute, University of York, York YO10 5DD, UK.

⁶Department for the Economics of Climate Change, Technical University Berlin, 10623 Berlin, Germany. ⁷Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

*To whom correspondence should be addressed. E-mail: glen.peters@cicero.uio.no

References

1. D. Guan, K. Hubacek, C. L. Weber, G. P. Peters, D. M. Reiner, *Glob. Environ. Change* **18**, 626 (2008).
2. G. P. Peters, C. L. Weber, D. Guan, K. Hubacek, *Environ. Sci. Technol.* **41**, 5939 (2007).
3. National Bureau of Statistics, *China Statistical Yearbook* (China Statistical Press, China, 1991–2008).
4. D. Guan, G. P. Peters, C. L. Weber, K. Hubacek, *Geophys. Res. Lett.* **36**, L04709 (2009).
5. C. L. Weber, G. P. Peters, D. Guan, K. Hubacek, *Energy Pol.* **36**, 3572 (2008).
6. G. P. Peters, J. Liu, E. G. Hertwich, R. Wang, paper presented at the Ninth Biennial Conference of the International Society for Ecological Economics on Ecological Sustainability and Human Well-Being, New Delhi, India, 15 to 18 December 2006.

Risks of Immune System Treatments

WE WISH TO ADD SEVERAL POINTS TO THE News Focus story “Replacing an immune system gone haywire” (J. Couzin-Frankel, 12 February, p. 772).

First, a great deal of research had already been done before the 1996 Basel meeting mentioned in the story. Stem cell transplants had been studied in animal models of autoimmune disease (1–5). Patient stem cell transplant protocols had been written, and a few human patients had already been treated specifically for autoimmune disease (6–10).

Second, we would like to stress the varying levels of risk in the treatment strategies described in the story. The immune system originates from hematopoietic stem cells (HSCs). Before receiving a transplant, patients with autoimmune diseases receive “conditioning” chemotherapy or radiation that destroys lymphocytes, inducing an immediate immune cease-fire. Subsequently, HSCs are infused to regenerate a new self-tolerant immune system. Sullivan and Nash advocate conditioning regimens with high doses of radiation. These extreme regimens cause irreversible bone marrow failure, thus requiring mandatory HSC reinfusion. The rationale for this high-dose strategy is that maximal ablation of the immune system will translate into longer and more durable disease remission. In contrast, we advocate less extreme regimens of chemotherapy, which can halt inflammation without altering the bone marrow’s ability to recover. The News Focus article also comments on the risk of infertility when patients are pretreated with chemotherapy. We emphasize that the risk of infertility is higher for the more extreme regimens.

TECHNICAL COMMENT ABSTRACTS

COMMENT ON “Detection of an Infectious Retrovirus, XMRV, in Blood Cells of Patients with Chronic Fatigue Syndrome”

Cathie Sudlow, Malcolm Macleod, Rustam Al-Shahi Salman, Jon Stone

Lombardi *et al.* (Reports, 23 October 2009, p. 585) reported an association between the human gammaretrovirus XMRV and chronic fatigue syndrome. However, their results may be misleading because of various potential sources of bias and confounding. If real, the association may lack generalizability because of the specific characteristics of the cases studied and could be due to reverse causality.

Full text at www.sciencemag.org/cgi/content/full/328/5980/825-a

COMMENT ON “Detection of an Infectious Retrovirus, XMRV, in Blood Cells of Patients with Chronic Fatigue Syndrome”

Andrew Lloyd, Peter White, Simon Wessely, Michael Sharpe, Dedra Buchwald

Lombardi *et al.* (Reports, 23 October 2009, p. 585) reported a significant association between the human retrovirus XMRV and chronic fatigue syndrome (CFS). However, the cases with CFS and the control subjects in their study are poorly described and unlikely to be representative. Independent replication is a critical first step before accepting the validity of this finding.

Full text at www.sciencemag.org/cgi/content/full/328/5980/825-b

COMMENT ON “Detection of an Infectious Retrovirus, XMRV, in Blood Cells of Patients with Chronic Fatigue Syndrome”

Jos W. M. van der Meer, Mihai G. Netea, Jochem M. D. Galama, Frank J. M. van Kuppeveld

Lombardi *et al.* (Reports, 23 October 2009, p. 585) reported detection of the human gammaretrovirus XMRV in the blood cells of patients with chronic fatigue syndrome (CFS). However, the patient description provided was incomplete. The inclusion of patients from a “CFS outbreak” previously linked with a viral infection, without confirmation in sporadic CFS cases, casts doubt on the role of XMRV in the pathogenesis of CFS.

Full text at www.sciencemag.org/cgi/content/full/328/5980/825-c

RESPONSE TO COMMENTS ON “Detection of an Infectious Retrovirus, XMRV, in Blood Cells of Patients with Chronic Fatigue Syndrome”

Judy A. Mikovits and Francis W. Ruscetti

We reported the detection of the human gammaretrovirus XMRV in 67% of 101 patients with chronic fatigue syndrome (CFS) and in 3.7% of 218 healthy controls, but we did not claim that XMRV causes CFS. Here, we explain why the criticisms of Sudlow *et al.*, Lloyd *et al.*, and van der Meer *et al.* regarding the selection of patients and controls in our study are unwarranted.

Full text at www.sciencemag.org/cgi/content/full/328/5980/825-d

Finally, although the News Focus story comments on problems obtaining insurance approval in the United States, medical funding is a worldwide issue, including in countries with government-funded health services. In addition to patient safety benefits, less toxic regimens also cost any health care system less money, because patients are less likely to suffer complications such as secondary cancers.

RICHARD K. BURT,^{1*} MARIO ABINUN,²

DOMINIQUE FARGE-BANCEL,³

ATHANASIOS FASSAS,⁴ FALK HIEPE,⁵

EVA HAVRDOVÁ,⁶ SUSUMU IKEHARA,⁷

YVONNE LOH,⁸ ALBERTO MARMONT DU HAUT

CHAMP,⁹ JÚLIO C. VOLTARELLI,¹⁰

JOHN SNOWDEN,¹¹ SHIMON SLAVIN¹²

¹Division of Immunotherapy, Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA. ²Newcastle General Hospital, Newcastle Upon Tyne, UK. ³Service de Médecine Interne et Pathologie Vasculaire, Inserm U976, Hôpital Saint-Louis, 75010 Paris, France. ⁴Hematology Department, Bone Marrow Transplantation Unit, Cell and Gene Therapy Center, George Papanicolaou Hospital, 57010 Exokhi, Thessaloniki, Greece. ⁵Charité University Medicine Berlin, Department of Rheumatology and Clinical Immunology and German Rheumatism Research Centre Berlin, A Leibniz Institute, Charitépaltz 1, D-10117 Berlin, Germany. ⁶MS Center, Department of Neurology, First Medical Faculty, Charles University in Prague, Katerinská 30, 128 08, Prague, Czech Republic. ⁷Department of Pathology, Regeneration Research Center

for Intractable Diseases, Center for Cancer Therapy, Kansai Medical University, 10-15 Fumizono-cho, Moriguchi City, Osaka 570-8506, Japan. ⁸Department of Haematology, HSCT Program, Singapore General Hospital, 169608 Singapore. ⁹Division of Hematology and Stem Cell Transplantation, Azienda-Ospedaliera Universitaria San Martino, Genoa, Italy. ¹⁰University of São Paulo (USP), Hemocentro Regional e Unidade de Transplante de Medula Óssea-HCF-MRP-USP, Rua Tenente Catão Roxo, 2501, CEP 14051-140, Ribeirão, Preto, São Paulo, Brazil. ¹¹Sheffield Teaching Hospitals NHS Trust, University of Sheffield, UK. ¹²The International Center for Cell Therapy and Cancer Immunotherapy (CTCI), Weizman Center, Tel Aviv, Israel.

*To whom correspondence should be addressed. E-mail: rburt@northwestern.edu

References

1. R. Yasumizu *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 6555 (1987).
2. S. Than *et al.*, *J. Exp. Med.* **176**, 1233 (1992).
3. D. M. Karussis *J. Clin. Invest.* **92**, 765 (1993).
4. M. van Gelder, D. W. van Bekkum, *Bone Marrow Transplant.* **18**, 1029 (1996).
5. R. K. Burt *et al.*, *J. Neurosci. Res.* **41**, 526 (1995).
6. A. Fassas *et al.*, *Bone Marrow Transplant.* **20**, 631 (1997).
7. A. M. Marmont, *Lupus* **2**, 151 (1993).
8. A. M. Marmont, M. T. van Lint, F. Gualandi, A. Bacigalupo, *Lupus* **6**, 545 (1997).
9. R. K. Burt, W. Burns, A. Hess, *Bone Marrow Transplant.* **16**, 1 (1995).
10. S. Slavin, *Bone Marrow Transplant.* **12**, 85 (1993).

CORRECTIONS AND CLARIFICATIONS

Letters: "Climate change and the integrity of science" by P. H. Gleick *et al.* (7 May, p. 689). Due to an editorial error, the original image was not a photograph but a collage. It was a mistake to have used it. The image (link available at www.sciencemag.org/cgi/content/full/328/5979/689/DC2) has been replaced in the HTML version and in the online PDF by an unaltered photograph from National Geographic (CREDIT: Paul Nicklen/National Geographic/Getty Images) of two polar bears on an ice floe.

News Focus: "Meeting briefs: The ins and outs of HIV" by J. Cohen (5 March, p. 1196). The earliest report of HIV predominantly entering cells through endocytosis appeared in C. D. Pauza, T. M. Price, *J. Cell Biol.* **107**, 959 (1988).

Letters to the Editor

Letters (~300 words) discuss material published in *Science* in the previous 3 months or issues of general interest. They can be submitted through the Web (www.submit2science.org) or by regular mail (1200 New York Ave., NW, Washington, DC 20005, USA). Letters are not acknowledged upon receipt, nor are authors generally consulted before publication. Whether published in full or in part, letters are subject to editing for clarity and space.



Summer Internships Students with Disabilities

AAAS started Entry Point!, a program that offers students with disabilities competitive internship opportunities in science, engineering, mathematics, computer science, and some fields of business. And this is just one of the ways that AAAS is committed to advancing science to support a healthy and prosperous world. Join us. Together we can make a difference.

To learn more, visit:
aaas.org/plusyou/entrypoint

 AAAS + U = Δ

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

The following resources related to this article are available online at www.sciencemag.org (this information is current as of May 13, 2010):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/328/5980/825-a>

This article **cites 6 articles**, 2 of which can be accessed for free:

<http://www.sciencemag.org/cgi/content/full/328/5980/825-a#otherarticles>

This article has been **cited by 1 articles** hosted by HighWire Press; see:

<http://www.sciencemag.org/cgi/content/full/328/5980/825-a#otherarticles>

This article appears in the following **subject collections**:

Virology

<http://www.sciencemag.org/cgi/collection/virology>

Comment on "Detection of an Infectious Retrovirus, XMRV, in Blood Cells of Patients with Chronic Fatigue Syndrome"

Cathie Sudlow,^{1,2*} Malcolm Macleod,^{1,3} Rustam Al-Shahi Salman,¹ Jon Stone¹

Lombardi *et al.* (Reports, 23 October 2009, p. 585) reported an association between the human gammaretrovirus XMRV and chronic fatigue syndrome. However, their results may be misleading because of various potential sources of bias and confounding. If real, the association may lack generalizability because of the specific characteristics of the cases studied and could be due to reverse causality.

Well-conducted case-control studies provide important insights into disease pathogenesis. Lombardi *et al.* (1) demonstrated an apparent association between chronic fatigue syndrome (CFS) and the presence, infectivity of, and immune response to the human gammaretrovirus, xenotropic murine leukemia virus-related virus (XMRV). However, their report has few features of a well-conducted case-control study, making their findings potentially misleading. Here, we briefly discuss several issues of study design, analysis, and interpretation, which may have contributed to inappropriate conclusions.

First, although the CFS cases studied fulfilled broadly accepted diagnostic criteria, they were selected from regions of CFS "outbreaks" and had specific immunological abnormalities, making them potentially more susceptible to viral infections than most patients with CFS. This could limit the generalizability of the study's results. Second, to avoid selection bias, the CFS-

free controls should have been drawn from the same background population as the cases and selected independent of the exposure (in this case, a viral infection) under study (2, 3). Put simply, the controls should ideally have been people who would have been cases in the study if they had CFS. However, the control subjects are not described in (1) beyond a mention that they were healthy donors. Third, the lack of clinical data for cases and controls makes it impossible to assess the potential for confounding by numerous other characteristics that may independently influence XMRV status, including age, sex, social deprivation status, medical history (e.g., of prostate cancer), and area of residence. Although confounding by demographic and clinical characteristics alone is unlikely to account for all of the observed difference in XMRV status between cases and controls, it could certainly be a contributory factor.

Fourth, Lombardi *et al.* do not explain whether identical and contemporaneous laboratory sample storage, handling, and analysis procedures were used for both cases and controls. Differences in these could be another potentially important source of confounding. Fifth, even if identical laboratory procedures for cases and controls were intended, researchers exploring an exciting new hypothesis of a viral cause for CFS in a laboratory established to explore biological causes

of CFS will be understandably eager for positive results. This so-called "expectation bias" may lead to completely unconscious and nondeliberate differences in sample handling and data interpretation between cases and controls; it can be avoided only if researchers are blinded to the case-control status of the samples. However, this is not described in (1).

Finally, the criteria for selecting samples for further analyses (including XMRV protein expression in 30/101 cases and 16/218 controls, infectivity in 12 cases and 12 controls, and immune response in 18 cases and 7 controls) are not described and are another important source of potential bias. Was this selection made at random or without knowledge of case-control status? Were all case and control samples on which testing was performed included in data analyses?

Aside from these crucial methodological issues, other plausible alternative explanations for the findings are not explicitly discussed. Foremost among these is reverse causality: Patients with poor general health because of CFS may be more susceptible to viral and other infections. We welcome biological research in CFS but are concerned that Lombardi *et al.* (1) have not discussed the possibilities of bias, confounding, reverse causality, and lack of generalizability in their study. Patients with CFS, understandably desperate for answers, deserve high-quality research into their condition. Full details of the methodological issues described above should be published, so that the scientific community can properly assess the credibility of these findings. This is particularly important because three studies (including a total of 388 patients with CFS and 438 controls from the United Kingdom and the Netherlands) have now failed to demonstrate any link between XMRV and CFS (4–6).

References and Notes

1. V. C. Lombardi *et al.*, *Science* **326**, 585 (2009).
2. K. F. Schulz, D. A. Grimes, *Lancet* **359**, 431 (2002).
3. D. A. Grimes, K. F. Schulz, *Lancet* **365**, 1429 (2005).
4. O. Erlwein *et al.*, *PLoS ONE* **5**, e8519 (2010).
5. H. C. Groom *et al.*, *Retrovirology* **7**, 10 (2010).
6. F. J. M. van Kuppeveld *et al.*, *BMJ* **340**, c1018 (2010).

19 October 2009; accepted 18 April 2010
10.1126/science.1183545

¹Division of Clinical Neurosciences, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK.

²Centre for Molecular Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK. ³Department of Neurology, NHS Forth Valley, Stirling, UK.

*To whom correspondence should be addressed. E-mail: cathie.sudlow@ed.ac.uk



LETTERS

edited by Jennifer Sills

China's Environmental Civilian Activism

IN THE POLICY FORUM "CHINA'S ROAD TO SUSTAINABILITY" (2 APRIL, P. 50), J. LIU OVERLOOKS AN important cultural force. China's worsening environmental conditions have catalyzed a spirit of environmental civilian activism.

For example, in 2003, a consortium proposed erecting 13 dams on the Nujiang River. China's environmental nongovernmental organizations and scholars launched a protest campaign through the Internet and newspapers. The critics argued that as reservoirs behind the dams filled up, flooding and landslides would imperil habitats. In response, Premier Wen Jiabao suspended the dam project pending an environmental review in 2004 (1).

The landmark of environmental civilian activism occurred in Xiamen City in 2007. The local government supported construction of a \$1.4-billion paraxylene plant near the center of the city. Information about the environmental impact of this project was not made available to the local residents. The people of Xiamen City were outraged when—through cell phone messages and the Internet—they learned of the plant's environmental risks. A phone text message was circulated among Xiamen citizens in late May calling for a "collective walk" (demonstration). On 1 June 2007, more than 1000 citizens gathered in front of the municipal building to protest. The demonstration forced the local government to cancel the largest industrial project in the history of Xiamen (2).



Protests. Chinese citizens protest against the planned Guangzhou trash incinerator in 2009. Their banners read, "Oppose the trash incinerator."

by the authorities. (In earlier years, standard practice was to obey Beijing-based experts in environmental protection.) In addition, the self-organized middle class forced the local government to open discussion by Internet. By seizing the opportunity for an open discussion, the newly empowered locals took to the streets to protect their environmental rights (4).

Recent years have witnessed an impressive growth in environmental protests in China. The number of petitions and mass public protests related to environmental issues has increased by 30% per year in the past few years, although the number of petitions lodged with the Chinese government has dropped (5).

The current environmental civilian activism movements have several common characteristics: (i) They are confined to one specific geographical space. (ii) Their goal is protecting the environment, rather than political rights or commercial interests. (iii) They focus on a specific pollutant, rather than general environmental degradation. The local nature of the movement enables the organization of a large number of citizens with little effort in a very short time. Given more open social and political conditions and the increasing size of the middle class in

The burgeoning middle class has become the driver of environmental civilian activism. For example, operation of the Likeng trash incinerator in Guangzhou City started in 2005 without any protest, although local farmers worried about health risk (3). In contrast, the proposed Panyu trash incinerator in Guangzhou City in 2009 triggered protests that were led by the middle class (4), who used science-based evidence to openly challenge prevailing notions formulated

China, environmental civilian activism will certainly be a key driver in China's transition to sustainability.

QIANG WANG

Qingdao Institute of Bioenergy and Bioprocess Technology, Chinese Academy of Science, Qingdao, 266101, China; Guangzhou Institute of Geochemistry, Chinese Academy of Science, Guangzhou, 510640, China. E-mail: tswangqiang@yahoo.com.cn

References and Notes

1. R. Stone, *Science* **319**, 1604 (2008).
2. F. Zeng, Z. Jiang, "Annual figures: Xiamen citizen in PX protest," *Southern People Weekly Magazine* (18 January 2008); www.infzm.com/content/4640 (in Chinese).
3. J. Chen, "Some farmers near Likeng trash incinerator died of cancer," *China News Weekly* (4 December 2009); reproduced at http://news.sina.com.cn/c/sd/2009-11-26/111519132701_2.shtml (in Chinese).
4. M. Moore, "China's middle-class rise up in environmental protest," *Telegraph* (23 November 2009); www.telegraph.co.uk/news/worldnews/asia/china/6636631/Chinas-middle-class-rise-up-in-environmental-protest.html.
5. S. Zhou, "Creating environmental supervision and environmental law enforcement to cope with new task and new challenges" (Ministry of Environmental Protection of the People's Republic of China, 2008); www.mep.gov.cn/info/ldjh/200804/t20080425_121679.htm (in Chinese).
6. The views expressed in this paper are the author's own and not necessarily those of QIBET-CAS, GIG-CAS.

I thank B. Jong for comments and linguistic support.

Effects of China's Economic Growth

IN A RECENT POLICY FORUM, J. LIU REVIEWED "China's road to sustainability" (2 April, p. 50). Liu focused on population growth and an increase in the number of households, but he failed to adequately address the most important socioeconomic driver behind environmental degradation in China: rapid economic growth that is not offset by efficiency improvements (1, 2).

In China, exports and capital investments contribute significantly more to gross domestic product (GDP) than household and government consumption combined (3), and this also holds true for emissions (1, 2). From 2002 to 2005, the production of exports was responsible for 50% of the growth in carbon dioxide emissions and capital formation was responsible for 35%;



Global warming
and lizards

832



Mobile small
RNA

834

household and government consumption contributed 15% (4, 5).

Liu discusses possible population control and household size, but a more dominant issue in terms of population dynamics is the migration from rural to urban areas (6). From 1990 to 2007, the urban population increased by 292 million, whereas the rural population decreased by 116 million (3). Urban dwellers, even if migrants from rural areas, have a higher income (3) and hence higher energy use and environmental impacts (2, 6).

A key challenge for China is to continue strong economic growth while minimizing environmental impacts. Reductions in emis-

sions per unit of GDP are unlikely to reduce total emissions if economic growth continues (1). China will need to combine aggressive domestic policies with international support to reverse the current growth in coal-dominated energy use and emissions.

GLEN P. PETERS,^{1*} DABO GUAN,²
KLAUS HUBACEK,^{3,4} JAN C. MINX,^{5,6}
CHRISTOPHER L. WEBER⁷

¹Center for International Climate and Environmental Research—Oslo (CICERO), Norway. ²Cambridge Centre for Climate Change Mitigation Research, Department of Land Economy, University of Cambridge, UK. ³Sustainability Research Institute, School of Earth and Environment, University of Leeds, UK. ⁴Department of Geography, University of Maryland, College Park, MD 20742, USA. ⁵Stockholm Environment Institute, University of York, York YO10 5DD, UK.

⁶Department for the Economics of Climate Change, Technical University Berlin, 10623 Berlin, Germany. ⁷Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

*To whom correspondence should be addressed. E-mail: glen.peters@cicero.uio.no

References

1. D. Guan, K. Hubacek, C. L. Weber, G. P. Peters, D. M. Reiner, *Glob. Environ. Change* **18**, 626 (2008).
2. G. P. Peters, C. L. Weber, D. Guan, K. Hubacek, *Environ. Sci. Technol.* **41**, 5939 (2007).
3. National Bureau of Statistics, *China Statistical Yearbook* (China Statistical Press, China, 1991–2008).
4. D. Guan, G. P. Peters, C. L. Weber, K. Hubacek, *Geophys. Res. Lett.* **36**, L04709 (2009).
5. C. L. Weber, G. P. Peters, D. Guan, K. Hubacek, *Energy Pol.* **36**, 3572 (2008).
6. G. P. Peters, J. Liu, E. G. Hertwich, R. Wang, paper presented at the Ninth Biennial Conference of the International Society for Ecological Economics on Ecological Sustainability and Human Well-Being, New Delhi, India, 15 to 18 December 2006.

Risks of Immune System Treatments

WE WISH TO ADD SEVERAL POINTS TO THE News Focus story “Replacing an immune system gone haywire” (J. Couzin-Frankel, 12 February, p. 772).

First, a great deal of research had already been done before the 1996 Basel meeting mentioned in the story. Stem cell transplants had been studied in animal models of autoimmune disease (1–5). Patient stem cell transplant protocols had been written, and a few human patients had already been treated specifically for autoimmune disease (6–10).

Second, we would like to stress the varying levels of risk in the treatment strategies described in the story. The immune system originates from hematopoietic stem cells (HSCs). Before receiving a transplant, patients with autoimmune diseases receive “conditioning” chemotherapy or radiation that destroys lymphocytes, inducing an immediate immune cease-fire. Subsequently, HSCs are infused to regenerate a new self-tolerant immune system. Sullivan and Nash advocate conditioning regimens with high doses of radiation. These extreme regimens cause irreversible bone marrow failure, thus requiring mandatory HSC reinfusion. The rationale for this high-dose strategy is that maximal ablation of the immune system will translate into longer and more durable disease remission. In contrast, we advocate less extreme regimens of chemotherapy, which can halt inflammation without altering the bone marrow’s ability to recover. The News Focus article also comments on the risk of infertility when patients are pretreated with chemotherapy. We emphasize that the risk of infertility is higher for the more extreme regimens.

TECHNICAL COMMENT ABSTRACTS

COMMENT ON “Detection of an Infectious Retrovirus, XMRV, in Blood Cells of Patients with Chronic Fatigue Syndrome”

Cathie Sudlow, Malcolm Macleod, Rustam Al-Shahi Salman, Jon Stone

Lombardi *et al.* (Reports, 23 October 2009, p. 585) reported an association between the human gammaretrovirus XMRV and chronic fatigue syndrome. However, their results may be misleading because of various potential sources of bias and confounding. If real, the association may lack generalizability because of the specific characteristics of the cases studied and could be due to reverse causality.

Full text at www.sciencemag.org/cgi/content/full/328/5980/825-a

COMMENT ON “Detection of an Infectious Retrovirus, XMRV, in Blood Cells of Patients with Chronic Fatigue Syndrome”

Andrew Lloyd, Peter White, Simon Wessely, Michael Sharpe, Dedra Buchwald

Lombardi *et al.* (Reports, 23 October 2009, p. 585) reported a significant association between the human retrovirus XMRV and chronic fatigue syndrome (CFS). However, the cases with CFS and the control subjects in their study are poorly described and unlikely to be representative. Independent replication is a critical first step before accepting the validity of this finding.

Full text at www.sciencemag.org/cgi/content/full/328/5980/825-b

COMMENT ON “Detection of an Infectious Retrovirus, XMRV, in Blood Cells of Patients with Chronic Fatigue Syndrome”

Jos W. M. van der Meer, Mihai G. Netea, Jochem M. D. Galama, Frank J. M. van Kuppeveld

Lombardi *et al.* (Reports, 23 October 2009, p. 585) reported detection of the human gammaretrovirus XMRV in the blood cells of patients with chronic fatigue syndrome (CFS). However, the patient description provided was incomplete. The inclusion of patients from a “CFS outbreak” previously linked with a viral infection, without confirmation in sporadic CFS cases, casts doubt on the role of XMRV in the pathogenesis of CFS.

Full text at www.sciencemag.org/cgi/content/full/328/5980/825-c

RESPONSE TO COMMENTS ON “Detection of an Infectious Retrovirus, XMRV, in Blood Cells of Patients with Chronic Fatigue Syndrome”

Judy A. Mikovits and Francis W. Ruscetti

We reported the detection of the human gammaretrovirus XMRV in 67% of 101 patients with chronic fatigue syndrome (CFS) and in 3.7% of 218 healthy controls, but we did not claim that XMRV causes CFS. Here, we explain why the criticisms of Sudlow *et al.*, Lloyd *et al.*, and van der Meer *et al.* regarding the selection of patients and controls in our study are unwarranted.

Full text at www.sciencemag.org/cgi/content/full/328/5980/825-d

Finally, although the News Focus story comments on problems obtaining insurance approval in the United States, medical funding is a worldwide issue, including in countries with government-funded health services. In addition to patient safety benefits, less toxic regimens also cost any health care system less money, because patients are less likely to suffer complications such as secondary cancers.

RICHARD K. BURT,^{1*} MARIO ABINUN,²

DOMINIQUE FARGE-BANCEL,³

ATHANASIOS FASSAS,⁴ FALK HIEPE,⁵

EVA HAVRDOVÁ,⁶ SUSUMU IKEHARA,⁷

YVONNE LOH,⁸ ALBERTO MARMONT DU HAUT

CHAMP,⁹ JÚLIO C. VOLTARELLI,¹⁰

JOHN SNOWDEN,¹¹ SHIMON SLAVIN¹²

¹Division of Immunotherapy, Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA. ²Newcastle General Hospital, Newcastle Upon Tyne, UK. ³Service de Médecine Interne et Pathologie Vasculaire, Inserm U976, Hôpital Saint-Louis, 75010 Paris, France. ⁴Hematology Department, Bone Marrow Transplantation Unit, Cell and Gene Therapy Center, George Papanicolaou Hospital, 57010 Exokhi, Thessaloniki, Greece. ⁵Charité University Medicine Berlin, Department of Rheumatology and Clinical Immunology and German Rheumatism Research Centre Berlin, A Leibniz Institute, Charitépaltz 1, D-10117 Berlin, Germany. ⁶MS Center, Department of Neurology, First Medical Faculty, Charles University in Prague, Katerinská 30, 128 08, Prague, Czech Republic. ⁷Department of Pathology, Regeneration Research Center

for Intractable Diseases, Center for Cancer Therapy, Kansai Medical University, 10-15 Fumizono-cho, Moriguchi City, Osaka 570-8506, Japan. ⁸Department of Haematology, HSCT Program, Singapore General Hospital, 169608 Singapore. ⁹Division of Hematology and Stem Cell Transplantation, Azienda-Ospedaliera Universitaria San Martino, Genoa, Italy. ¹⁰University of São Paulo (USP), Hemocentro Regional e Unidade de Transplante de Medula Óssea-HCF-MRP-USP, Rua Tenente Catão Roxo, 2501, CEP 14051-140, Ribeirão, Preto, São Paulo, Brazil. ¹¹Sheffield Teaching Hospitals NHS Trust, University of Sheffield, UK. ¹²The International Center for Cell Therapy and Cancer Immunotherapy (CTCI), Weizman Center, Tel Aviv, Israel.

*To whom correspondence should be addressed. E-mail: rburt@northwestern.edu

References

1. R. Yasumizu *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 6555 (1987).
2. S. Than *et al.*, *J. Exp. Med.* **176**, 1233 (1992).
3. D. M. Karussis *J. Clin. Invest.* **92**, 765 (1993).
4. M. van Gelder, D. W. van Bekkum, *Bone Marrow Transplant.* **18**, 1029 (1996).
5. R. K. Burt *et al.*, *J. Neurosci. Res.* **41**, 526 (1995).
6. A. Fassas *et al.*, *Bone Marrow Transplant.* **20**, 631 (1997).
7. A. M. Marmont, *Lupus* **2**, 151 (1993).
8. A. M. Marmont, M. T. van Lint, F. Gualandi, A. Bacigalupo, *Lupus* **6**, 545 (1997).
9. R. K. Burt, W. Burns, A. Hess, *Bone Marrow Transplant.* **16**, 1 (1995).
10. S. Slavin, *Bone Marrow Transplant.* **12**, 85 (1993).

CORRECTIONS AND CLARIFICATIONS

Letters: "Climate change and the integrity of science" by P. H. Gleick *et al.* (7 May, p. 689). Due to an editorial error, the original image was not a photograph but a collage. It was a mistake to have used it. The image (link available at www.sciencemag.org/cgi/content/full/328/5979/689/DC2) has been replaced in the HTML version and in the online PDF by an unaltered photograph from National Geographic (CREDIT: Paul Nicklen/National Geographic/Getty Images) of two polar bears on an ice floe.

News Focus: "Meeting briefs: The ins and outs of HIV" by J. Cohen (5 March, p. 1196). The earliest report of HIV predominantly entering cells through endocytosis appeared in C. D. Pauza, T. M. Price, *J. Cell Biol.* **107**, 959 (1988).

Letters to the Editor

Letters (~300 words) discuss material published in *Science* in the previous 3 months or issues of general interest. They can be submitted through the Web (www.submit2science.org) or by regular mail (1200 New York Ave., NW, Washington, DC 20005, USA). Letters are not acknowledged upon receipt, nor are authors generally consulted before publication. Whether published in full or in part, letters are subject to editing for clarity and space.



AAAS is here.

Summer Internships Students with Disabilities

AAAS started Entry Point!, a program that offers students with disabilities competitive internship opportunities in science, engineering, mathematics, computer science, and some fields of business. And this is just one of the ways that AAAS is committed to advancing science to support a healthy and prosperous world. Join us. Together we can make a difference.

To learn more, visit:
aaas.org/plusyou/entrypoint

 AAAS + U = Δ

ENVIRONMENTAL HISTORY

Dam Controversy

Bill Luckin

The English Lake District holds a cherished place in the affections of huge numbers of national and international visitors. As much a cultural and literary construct as a pocket-sized region of unparalleled natural beauty, it epitomizes what is thought to comprise the unchanging in an ever more technological age. All the more so for northwestern Britons. The Lakes are located within an hour and a half's drive from the cradle of mercantile and industrial capitalism, centered on the great cities of Liverpool and Manchester, and the cluster of large and medium-size satellite communities that have rendered the region the most densely populated and motorway-ridden part of the United Kingdom. Britain, and particularly England—and even more so what Victorians christened the “new industrial districts”—have had good reason to protect indisputably nonurban space.

The alternative might have been an ever-expanding factory culture, or, later, as has too often been the case, burgeoning suburban fill-in between chunks of postindustrial wreckage. In a majority of the key northwestern centers themselves, the contemporary tasks of regeneration and greening have not been fully confronted. The Lake District, north Wales, and the Yorkshire Dales offer sanctuary. But too little attention has been given to large-scale investment in the public sphere and the regeneration of postindustrial communities that reproduce social, economic, and cultural inequality. As Bill Bryson memorably and glumly reported, “I arrived in Liverpool and they were having a litter festival” (1). Much remains to be done, but under the newly elected government, the watchword will be a reduction in public expenditure. The future looks unpromising.

Problems were different 150 years ago, although then, as now, complex variants of the town-country issue loomed large. These form one of the threads of

Harriet Ritvo's *The Dawn of Green*, a penetrating microstudy that mixes environmental, scientific, urban, and political history. Ritvo (a historian at the Massachusetts Institute of Technology) chronicles water-starved, late-19th-century Manchester's determination to

convert tiny Thirlmere—north of imposing Windermere, south of gloriously mysterious Derwentwater—into the world's largest reservoir. In the massive “shock city” of early industrialism itself, urgent action was pondered and planned

by a progressive, though financially canny, council; a long-serving and dynamic town clerk; a clear-sighted Waterworks Committee; and a nationally eminent though controversial engineer, John Bateman. Not surprisingly, environmentally motivated opponents of the scheme were inspired by the spirit of William Wordsworth. Would-be revolutionary turned intensely conservative poet laureate, Wordsworth had protested the incursion of the railway into the sleepy market town of Keswick. Drunken day-trippers, he claimed, would devastate lake, fell, and forest. They should stay at home, sustain their own communities, and devote sparse spare time to self-education and moral improvement. The great poet and his bohemian-aesthete friends and relations—an in-turned and sexually and

morally troubled lot—played a key role in inventing the exclusive and excluding idea of elite landscape heritage in the Lakes.

Heaven knows what they would have made of the conversion of rugged little Thirlmere into a low-walled and mock-castellated memorial to Manchester's unstoppable industrial development and (even more galling to the city's urban rivals, who believed they too had a claim on Lakeland water) overweening sense of civic self-importance. Protesters believed (although this may have been a rural legend) that in a solemn act of nature-affirming romantic togetherness, Wordsworth and his group had carved their names on a rock near the south end of Thirlmere itself.

“Philistine” and utilitarian Manchester wasn't to be outdone. After the laying of the foundation stone for the reservoir's embankment in 1890, the mayor, the Waterworks Committee, and all other members of the council commemorated themselves with a notice board stylistically identical to those that warned Victorians what they could and couldn't do in public parks. (Ritvo's photograph of this plaque conveys the curious deadness of Thirlmere compared with the magical aliveness of the smaller, wilder Lakes.)

The opposition had been divided. Wordsworth's successors, notably the aging Thomas Carlyle and the splenetic and at times half-mad John Ruskin, railed against the project, but in a style that indicated they had failed to understand the main points at issue. Landowners took a different line and held out for the highest price per acre that Manchester could be persuaded to pay. Naturalists worried about the fate of trout, grayling, pike, and char. But cash-strapped baronets, embarrassed by regional depression during the final

30 years of the 19th century, bowed to the inevitable (although, again, at a more than acceptable price). Farmers haggled over rights of way, then gave in when the money was right. The motley Thirlmere Defence Association lacked the singularity of purpose and activist energy that stymies or gains invaluable concessions from powerful public authorities when major and socially destabilizing utility, housing, or industrial projects come on line.

“The dawn of green environmentalism”? In

The Dawn of Green
Manchester, Thirlmere, and
Modern Environmentalism
by Harriet Ritvo
University of Chicago Press,
Chicago, 2009. 245 pp., \$26, £18.
ISBN 9780226720821.



Dammed waters. Photochrom print of Helvellyn and Thirlmere, Lake District, from the 1890s.

The reviewer is at the Department of History, University of Bolton, Bolton BL3 5AB, UK. E-mail: b.luckin@bolton.ac.uk

one sense, yes. The affluent and socially powerful did very nicely, thank you. Idealistic protectors of “nature,” heritage, and the status quo made little impact. By the later 20th century, Lakeland would be confronted by an altogether more serious problem: how could this pocket-sized paradise survive incessant tourist overkill? Perhaps Wordsworth hadn’t been entirely wrong about Keswick.

References

1. B. Bryson, *Notes from a Small Island* (Doubleday, London, 1995).

10.1126/science.1189236

ENERGY RESOURCES

Oil Peak or Panic?

David Lloyd Greene

Peak oil, a serious issue, is not about running out of oil. It is about rates: the rates of oil discovery and production, the rate at which demand grows, and the rate of technological change. The oil peakers’ contribution to understanding the world oil situation can be summed up as follows: rates matter as much or more than quantities, and geology matters as much or more than economics and technology. It is easy to caricature the oil peakers’ assessment as a mechanistic calculation about using up a fixed resource. It is also easy to caricature their opponents’ view as blind faith that markets and technology will overcome all problems. One of the things that makes Steven Gorelick’s *Oil*

Panic and the Global Crisis well worth reading is that he does neither. It is a book serious students of the world oil market should read, not because Gorelick has all the answers but because his account is well reasoned, well informed, and argued honestly, with respect for responsible opposing viewpoints.

Proponents of peak oil are sometimes referred to as “pessimists,” other times as “geologists.” Opponents are often called “optimists” or “economists.” Gorelick (a hydrogeologist at Stanford University) is clearly in the optimists’ camp, but not entirely.

The reviewer is at the Energy and Transportation Science Division, National Transportation Research Center, Oak Ridge National Laboratory, 2360 Chertala Boulevard, Knoxville, TN 37932, USA. E-mail: dlgreene@ornl.gov



To keep the oil flowing. Floating rigs allow drilling and production in deeper waters.

This can be inferred by examining the table of contents: He devotes 26 pages to “The Historical Resource Depletion Debate,” whereas he gives 107 pages to “Counter-Arguments to Imminent Global Oil Depletion.” In the latter chapter, he provides ample historical evidence that despite the finiteness of global mineral resources, they tend either to become less scarce over time or to become obsolete. Declining rates of new oil discovery have been offset by an increasing ability to find more oil in existing reservoirs and to recover a greater fraction of the oil in place.

The vast majority of the counter-arguments countering oil peaking are well supported. But in my opinion, Gorelick is too dismissive of M. King Hubbert’s achievement. He characterizes as “not very accurate” Hubbert’s 1956 prediction of a 1965 peak in oil production in the coterminous United States because the peak did not occur until 1970. In 1956, U.S. oil production had been on a generally increasing path for over half a century.

Given this trend, Hubbert’s prediction has to be judged substantially correct (which is not the case for his prediction of a global peak). It remains essentially correct despite the fact that his estimate of the total resource endowment was low, despite dramatic progress in the technology of finding and producing oil, despite far higher oil prices, and despite the discoveries of important new oil fields. Perhaps the prediction was a lucky guess, but I think there’s more than luck involved. It just isn’t feasible to continue increasing the rate at which conventional oil is produced until the last drop is pumped. Therefore, there must be a turning point.

But what is oil anyway? And what is a resource? Not surprisingly, much of the

argument over oil depletion is a result of how these terms are defined. Oil peakers focus on conventional oil. But, as Gorelick points out, there are relatively large quantities of unconventional fossil resources that can be converted to liquid hydrocarbon fuels at costs the world’s economy has demonstrated that it is willing to pay. Some of the costs cited in the book are overly optimistic: for example, petroleum from U.S. oil shale at \$22 to \$38 per barrel. Yet, the point is correct and well substantiated.

Nonetheless, expanding unconventional oil production won’t be easy and is probably undesirable. Energy companies find the development of unconventional resources a risky proposition for three reasons: high capital investment requirements, higher greenhouse gas emissions than conventional petroleum, and uncertainty about future oil prices. At what rate could or should the world expand production from unconventional sources?

Gorelick points out that oil production has already peaked in many regions of the world: “The U.S. is the largest oil-producing nation that has experienced peak oil production, but other countries have also followed a pattern of decline.... British Petroleum reports that there are at least 25 countries producing oil below their peak values by 20 percent or more.”

Given this, it is disappointing that the book does not explore in greater detail the implications for world oil supply of the growing number of declining regions. The International Energy Agency has studied this subject in depth and predicted not a peak but a plateau in non-Organization of the Petroleum Exporting Countries (OPEC) production, right about now. Not very different projections have been produced by ExxonMobil. A plateau in non-OPEC production implies increasing dependence on OPEC, a massive transition to high-carbon unconventional fossil resources, higher and more volatile oil prices, a transition to electricity or hydrogen guided by public policy, or a combination of the above. The timing, extent, and intensity of oil peaking will probably strongly influence whether the transitions are relatively easy or painful. *Oil Panic* demonstrates convincingly that in the long run there will be more oil and replacements for oil. But as John Maynard Keynes observed, “In the long run, we are all dead.” The transition from oil matters, and we need to understand it better.

10.1126/science.1189537

Oil Panic and the Global Crisis Predictions and Myths

by Steven M. Gorelick

Wiley-Blackwell, Oxford, 2010.
255 pp., illus. \$54.95, £32.50.
ISBN 9781405195485.

RESEARCH ETHICS

Beyond Access vs. Protection in Trials of Innovative Therapies

Alex John London,¹ Jonathan Kimmelman,^{2*} Marina Elena Emborg³

Review of first-in-human trials should safeguard the integrity of the scientific enterprise through a focus on preclinical and clinical study quality.

Over the past decade, researchers have initiated innovative, early-phase, clinical trials, including the first-ever testing in humans of cell therapies for myocardial infarction (1), as well as transplantation of embryo-derived tissues for spinal cord repair (2). The next decade promises more initiatives involving first-in-human trials of innovative therapeutic strategies (3).

Such studies frequently inspire ethical debates that revolve around the rights and welfare of research participants. However, neither “protection” nor “access” (discussed below) adequately addresses issues relating to the methodological quality of preclinical (i.e., nonhuman) studies, the design and review of early-phase clinical studies, and the continuity between preclinical studies and first-in-human trials. These issues bear crucially on the value of a translational trial as a critical stage in a larger, collaborative endeavor with a unique social mission: transforming hard-won advances in basic science into practical applications that improve care at the bedside. Before the welfare and autonomy of participants comes into view, issues relating to the ability of science to fulfill this social mission—what we hereafter refer to as the integrity of the scientific enterprise—must be adequately addressed.

Protection and Access

The autonomy and welfare of study participants is, and ought to remain, a central concern for research ethics. Scientists and reviewers charged with protecting subjects dwell on uncertainties and risk. They urge cautious study designs and extended preclinical testing, often warning that volunteer patients who have exhausted standard care options are prone to overestimating therapeutic benefit (“protection”) (4). However, patient advocates and clinical investigators often argue that requests for additional preclinical stud-

ies—which, if conducted, delay translational trials (small trials of therapies emerging from laboratory studies)—do not advance the interests of people with untreatable diseases (“access”) (5). If we respect people by allowing them to choose in accordance with their values and to shepherd their interests as they see fit, then preventing patients from pursuing what they may regard as their best therapeutic option disadvantages participants and slows scientific progress.

However, exclusive focus on personal interests of subjects fails to assign proper weight to a range of ethical issues that arise in clinical research. Medical research is primarily directed toward producing a common, rather than a private, good: It serves an inherently social purpose of generating knowledge requisite for institutions to better address unmet health needs of community members (6). It is also a deeply social activity (7) in that new knowledge and interventions are produced from a long chain of investigations, each building on the last. Yet each link in this chain is also a discrete interaction between specific stakeholders with their own individual interests (see figure, above). Because uncoordinated activities of individuals pursuing personal interests can have deleterious effects on attainment and preservation of social goods, research ethics must place more emphasis on norms that preserve relationships and institutions necessary to sustain this social good.

Individual Transactions and Social Goods

Although individual trials, each a transaction between investigators and trial participants, are regulated by ethics committees and drug regulatory authorities, there are at least three ways they can inadvertently undermine forms of broader, longer-term collaboration that sustain the production of socially valuable medical knowledge. First, studies can misallocate resources. Practically every clinical trial makes demands on social and material resources beyond those provided by study sponsors, investigators, and human subjects. For example, clinical trials deplete the pool of



Continuity and quality. Innovative clinical research requires the sustained cooperation of diverse stakeholders, including patients and physician-researchers (as seen here), as well as scientists, sponsors, and institutions.

eligible volunteers for other meritorious trials. About 20% of trials initiated in National Cancer Institute (NCI)-designated Comprehensive Cancer Centers fail to accrue a single study volunteer (8), and fewer than 60% of NCI-funded clinical trials are able to meet minimal recruitment goals (9). Also, many early-phase trials are pursued at research centers that are heavily subsidized by public funding agencies and private philanthropies, and their management and oversight draw heavily on limited administrative resources. Poorly justified trials may also compete with better ones for highly specialized expertise and equipment. Finally, because they generally do not indemnify patients in the event of trial-related injuries, studies exact demands on third-party payers. The determination of whether a research protocol has sufficient prospect of returning social value to warrant such use of human and material resources has to be made in light of factors beyond the personal interests of researchers and potential trial participants (10).

Second, adversities encountered in isolated trials can have cascading effects, undermining institutional and social supports for new initiatives. Research programs are joined or abandoned on the basis of perceived beliefs of others (11). Unsuccessful or poorly conceived trials can dampen interest in promising therapies. Repeated failures can diminish the standing of a research program, interrupting recruitment of talent, investment, and

¹Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213, USA. ²Biomedical Ethics Unit, Social Studies of Medicine, McGill University, Montreal, Quebec H3A 1X1, Canada. ³Wisconsin National Primate Research Center and Department of Medical Physics, University of Wisconsin, Madison, WI 53705, USA.

*Author for correspondence. E-mail: jonathan.kimmelman@mcgill.ca

institutional support. Major adverse events can undermine support for an entire field. For example, the unexpected death of a volunteer in a gene-transfer study seriously eroded confidence in the field's ability to self-regulate; many suggest that this led to a withdrawal of investment and institutional support (12–14).

Third, oversight or review procedures that allow the preferences of researchers and participants to dominate the design and conduct of translational trials can create incentives that reward low-quality trials and penalize well-designed protocols. For example, deficiencies in a trial that affect the value of the information it generates may not discourage fully informed volunteer patients having few other care options. However, clinical trial protocols are examples of what economists call “credence goods”: Products whose quality is difficult to judge by consumers. Unless well-regulated, markets for credence goods are prone to low-quality products, because consumers lack the capacity to reward producers of high-quality products and to punish producers of mediocre ones (15, 16). Oversight structures that establish baseline quality for translational trials ensure that participants, physicians, researchers, and investors can pursue their individual interests without compromising the social mission of the research enterprise. Volunteer subjects can be confident that the trial is based on promising, well-designed protocols; all other stakeholders have greater assurance that the program represents an efficient means of producing a social good.

The Need for Quality Preclinical Data

Elsewhere, we have argued that rigorously designed preclinical studies greatly enhance the interpretability of translational trials (e.g., because they enable researchers to troubleshoot interventions when clinical outcomes are discordant with those in animal models) (13, 17). But there is a growing literature documenting quality deficiencies in preclinical research used to support phase 1 trials (18). In contrast with clinical research, preclinical research has only sporadically taken up measures to control bias, including a priori statement of hypothesis, random treatment allocation, blinded outcome assessment, and accounting for missing data (19, 20). The preclinical literature shows evidence of publication bias (19, 21), and not infrequently, first-in-human trials are initiated before preclinical studies have been subjected to peer review. Even when preclinical studies are rigorously designed and executed, first-in-human clinical studies may deviate in significant ways from the methods or procedures evaluated in preclinical studies.

Safeguarding the integrity of the scientific enterprise by ensuring that decisions about clinical trials are based on high-quality preclinical information thus entails resolving four key questions at the inception of a translational trial. First, do preclinical experiments provide a credible measure of effect? Are studies internally valid and have they been replicated independently? Second, do preclinical studies have reasonable external validity? Have investigators demonstrated that their preclinical findings are robust and generalizable and that their choice of animal models and outcome measures are justified? Third, to what extent have researchers justified the assumption that observations in a preclinical system will be reproduced in human patients? Have they presented data showing that causal preconditions of effect in animals are also present in human beings, and have they made a thorough search of the evidence? Last, when all the previous issues are addressed, do conditions used in preclinical studies correspond with those in a proposed human trial? Are delivery strategies, targets, doses, and materials in the human study identical or substantially equivalent to those validated in animals?

Ensuring the Integrity of the Enterprise

The primary responsibility for addressing these questions lies with preclinical and clinical investigators pursuing translational research. These questions should also be a focus of the various levels of ethical and scientific review that occur at each stage of the translational research process. This includes oversight by an Institutional Review Board (IRB) and committees that provide centralized review of protocols, such as the Recombinant DNA Advisory Committee and Data Monitoring Committees. These questions may also help to sensitize sponsors of translational trials to the ethical aspects of basic and preclinical research, not only when evaluating funding requests, but in ensuring that sufficient resources are in place to address them. Finally, these issues should be central in peer review and editorial scrutiny of translational studies.

IRBs may be reluctant to address these questions because such methodological issues may seem beyond their purview. But the primary value of first-in-human trials lies in the information they are expected to generate. IRBs must evaluate the quality of that information and its potential social value as part of the process of ensuring that risks are reasonable. It might be objected that these concerns run afoul of the statement in U.S. federal regulations that IRBs “should not consider possible long-range effects of applying knowledge gained in the

research (for example, the possible effects of the research on public policy)” (22). But this is meant to prevent scientifically valuable research from being stifled because of how sensitive or controversial findings might be used at a social level. It does not prohibit IRBs from scrutinizing the quality of the science that is likely to emerge from an investigation.

The dichotomy of subject protection versus access is insufficiently sensitive to ways in which clinical trials contribute to, and depend crucially upon, the quality of other scientific investigations. This is most vividly illustrated by first-in-human trials that involve members of a disease population. However, such issues arise in any form of research that draws heavily on sustained collaboration and coordination across time of diverse stakeholders. A focus on the integrity of the research enterprise draws attention to methodological and social requirements that medical science must satisfy if it is to maintain the support of those whose cooperation makes it possible and whose interests it is supposed to serve.

References and Notes

1. J. Couzin, G. Vogel, *Science* **304**, 192 (2004).
2. J. Alper, *Nat. Biotechnol.* **27**, 213 (2009).
3. California Institute for Regenerative Medicine, press release; www.cirm.ca.gov/PressRelease_102809.
4. B. Lo et al., *Stem Cells* **23**, 1454 (2005).
5. *Lancet* **365**, 1984 (2005).
6. A. J. London, *J. Med. Philos.* **32**, 99 (2007).
7. M. Dixon-Woods, C. Tarrant, *Soc. Sci. Med.* **68**, 2215 (2009).
8. D. M. Dilts, A. B. Sandler, *J. Clin. Oncol.* **24**, 4545 (2006).
9. G. A. Curt, B. A. Chabner, *Oncologist* **13**, 923 (2008).
10. E. J. Emanuel, D. Wendler, C. Grady, *JAMA* **283**, 2701 (2000).
11. A. Tatsioni, N. G. Bonitsis, J. P. Ioannidis, *JAMA* **298**, 2517 (2007).
12. Even well-designed studies can go awry or cause major adverse events. Our point is that adverse outcomes from poorly designed studies are more likely to cause negative cascades because the revelation of avoidable shortcomings is likely to cause recriminations that undermine the trust or support of stakeholders.
13. J. Kimmelman, *Gene Transfer and the Ethics of First-In-Human Experiments: Lost in Translation* (Cambridge Univ. Press, New York, 2010).
14. J. M. Wilson, *Science* **324**, 727 (2009).
15. M. T. Law, “History of food and drug regulation in the United States,” EH.Net, R. Whaples, Ed. (2004); <http://eh.net/encyclopedia/article/Law.Food.and.Drug.Regulation>.
16. D. Carpenter, in *Government and Markets: Toward a New Theory of Regulation*, E. Balleisen and D. Moss, Eds. (Cambridge Univ. Press, New York, 2009), pp. 164–190.
17. J. Kimmelman et al., *Mov. Disord.* **24**, 1893 (2009).
18. M. R. Macleod et al., *Stroke* **40**, e50 (2009).
19. P. Perel et al., *BMJ* **334**, 197 (2007).
20. V. Bebar, D. Luyten, K. Heard, *Acad. Emerg. Med.* **10**, 684 (2003).
21. M. Benatar, *Neurobiol. Dis.* **26**, 1 (2007).
22. Code of Federal Regulation 46.111(a)(2).
23. J.K. was funded by the Canadian Institutes of Health Research (NRF 80045 and EOG 102824). M.E.E. thanks Kinetics Foundation and NIH (P51RR000167) for support.

10.1126/science.1189369

BIOCHEMISTRY

Assembling the Outer Membrane

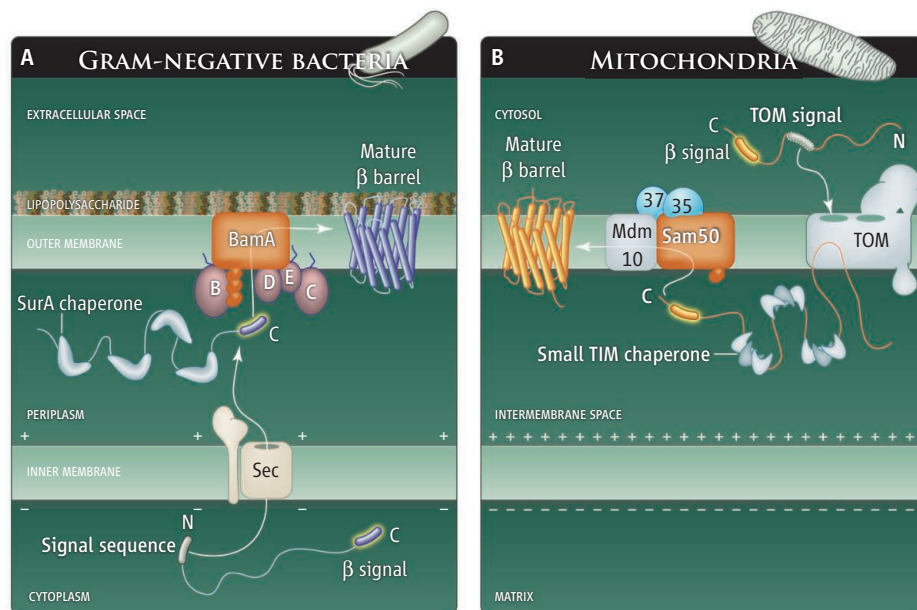
David A. Stroud, Chris Meisinger, Nikolaus Pfanner, Nils Wiedemann

Transmembrane proteins with β -barrel structure are characteristic of the outer membranes of Gram-negative bacteria, mitochondria, and chloroplasts (1). β -Barrel membrane proteins typically include a central pore that is lined by multiple antiparallel β strands. The functions of these proteins range from protein transport to metabolite flux. On page 890 of this issue, Hagan *et al.* (2) report a major step toward a molecular understanding of β -barrel membrane insertion and folding.

The machineries that mediate insertion and assembly of β -barrel proteins into the outer membrane are termed β -barrel assembly machinery (BAM) in bacteria (3, 4) and sorting and assembly machinery (SAM) in mitochondria (5–7). Most BAM and SAM components are not related to each other, but the central component (BamA and Sam50, respectively) has been conserved from bacteria to mitochondria. By reconstituting the *Escherichia coli* BAM complex in vitro, Hagan *et al.* now show that this sophisticated machine functions without an external supply of energy.

The central component of the bacterial machinery for β -barrel assembly is the transmembrane protein BamA (3). Four accessory lipoproteins exposed to the periplasm (BamB, BamC, BamD, and BamE) are associated with BamA, forming the BAM complex (see the figure, panel A) (2, 4). The precursors of outer-membrane β -barrel proteins are synthesized in the bacterial cytoplasm. The unfolded precursor proteins are recognized and translocated across the inner membrane by the Sec translocase. On arrival in the periplasm and proteolytic removal of the amino-terminal signal sequence, β -barrel precursors are protected from aggregation and misfolding by soluble chaperones (SurA, Skp, and DegP). These chaperones function in two discrete pathways, one involving SurA, the other Skp and DegP, but SurA alone delivers most of the substrates to the BAM in the outer membrane (8).

A key step for the mechanistic analysis of a complex biochemical process is the establishment of an in vitro system that dissects the process into distinct stages. β -Barrel



Outer-membrane biogenesis. The BAM complex in Gram-negative bacteria and the SAM complex in mitochondria catalyze the folding and membrane insertion of outer-membrane β -barrel proteins. (A) In bacteria, precursor proteins with an amino-terminal signal sequence are transported across the inner membrane by the Sec translocase. Periplasmic chaperones (SurA) maintain solubility and prevent aggregation of the precursor proteins. Targeting to the BAM complex is achieved by the β signal. (B) In mitochondria, precursor proteins synthesized in the cytosol are translocated across the outer membrane by the TOM complex (the nature of the TOM signal is so far unknown). Small TIM chaperones guide the precursors to the SAM complex, where they are recognized via the β signal.

transport and assembly have been studied in isolated mitochondria (5, 7, 9–11), but bacterial β -barrel assembly has been analyzed in vivo (3, 4, 8). Hagan *et al.* have not only set up an in vitro assembly system, but they also defined the minimal system required for β -barrel assembly. The authors first expressed BamAB and BamCDE subcomplexes in *E. coli* cells, then reconstituted the full BAM complex in vitro and incorporated it into artificial lipid vesicles (liposomes). To monitor β -barrel assembly, the authors used the precursor of an outer-membrane protease as substrate. Before incubation with liposomes containing the reconstituted BAM complex, a preincubation step between the unfolded precursor protein and the periplasmic chaperone SurA was required. Membrane insertion and folding of the substrate protein did not require an added external energy source.

Does the reconstitution of the bacterial machinery help in understanding the assembly machineries for β -barrel proteins of mitochondria and chloroplasts? In comparison to the bacterial system, the eukaryotic precursor proteins are synthesized on the opposite

Multisubunit machineries insert β -barrel proteins into the outer membranes of bacteria and mitochondria.

side of the outer membrane (5–7, 9–12). The precursors are first translocated across the mitochondrial outer membrane by the TOM complex (see the figure, panel B). In the intermembrane space that corresponds to the bacterial periplasm, they are bound by chaperone complexes formed by small TIM proteins (6). The topological position of the bound precursor proteins is similar to that in bacteria and the proteins can access the conserved core of the assembly machinery, Sam50, from the same side of the membrane as that from which BamA is accessed in bacteria. The further subunits of the SAM and BAM complexes are not related to each other, but the basic mechanism of membrane insertion is conserved, given that at least some bacterial β -barrel proteins can be handled by the mitochondrial machinery (13). The identification of the sorting signals that direct membrane insertion of mitochondrial and bacterial precursors through SAM and BAM supports a related assembly mechanism (see the figure) (9, 14).

The reconstituted system established by Hagan *et al.* will be invaluable for defining the

Institut für Biochemie und Molekularbiologie, ZBMZ, Trinationales Graduiertenkolleg 1478, Fakultät für Biologie, and Centre for Biological Signalling Studies, Universität Freiburg, 79104 Freiburg, Germany. E-mail: nikolaus.pfanner@biochemie.uni-freiburg.de

molecular functions of individual components of the assembly machinery. So far, three major mechanisms have been proposed for insertion of the β -barrel precursors into the outer membrane. BamA and Sam50, which are β -barrel proteins themselves, form channels (9, 14), but it is not yet clear whether the precursors are inserted into a pore formed within a monomer or whether oligomeric forms of BamA and Sam50 form a channel. A third possibility is that the precursor proteins are not inserted into the outer membrane via a protein channel, but that BAM/SAM function as a scaffold that facilitates insertion of β -barrel proteins at the protein-lipid interface.

What are the functions of the further components of the β -barrel assembly pathway? When Hagan *et al.* altered the composition of the accessory lipoproteins in their reconstituted system; they found that most subcomplexes were unstable. The stability of two subcomplexes—BamAB and BamACDE—was similar to that of the full BAM complex, but they had low activity in β -barrel assembly upon reconstitution into liposomes. Thus, all four accessory lipoproteins are required for full activity of the BAM complex. The results of Hagan *et al.* further suggest that multiple

copies of SurA bind to a precursor protein and that the chaperone-precursor complex directly delivers the β -barrel polypeptides to the outer-membrane assembly machinery in a folding-competent state (8). The authors found no evidence for an external energy source driving precursor transfer and insertion into the outer membrane. The transport pathway may be driven by the free energy released during folding and insertion of the β barrels into the lipid phase of the outer membrane.

The core processes of β -barrel biogenesis have been conserved during evolution (1), but the machineries acquired further functions. The mitochondrial SAM complex is not only required for the biogenesis of β -barrel proteins, but is also a dynamic platform for the assembly of α -helical proteins of the outer membrane. Several forms of the SAM complex, differing in subunit composition, serve distinct functions in the biogenesis of different classes of precursor proteins (12). The SAM complex is associated with a multifunctional organizing center that is involved in lipid transport, maintenance of mitochondrial shape, and the connection of mitochondria to the endoplasmic reticulum (10–12, 15). Future studies will address whether the BAM

complex is dedicated to β -barrel assembly only or whether it may play further roles in outer-membrane biogenesis. The reconstruction of the BAM complex as a stable, active, and homogeneous complex by Hagan *et al.* will also greatly aid in solving the high-resolution structure of an outer-membrane assembly machinery.

References

1. E. Schleiff, J. Soll, *EMBO Rep.* **6**, 1023 (2005).
2. C. L. Hagan *et al.*, *Science* **328**, 890 (2010); published online 8 April 2010 (10.1126/science.1188919).
3. R. Voulhoux, M. P. Bos, J. Geurtsen, M. Mols, J. Tommassen, *Science* **299**, 262 (2003).
4. T. Wu *et al.*, *Cell* **121**, 235 (2005).
5. N. Wiedemann *et al.*, *Nature* **424**, 565 (2003).
6. W. Neupert, J. M. Herrmann, *Annu. Rev. Biochem.* **76**, 723 (2007).
7. N. C. Chan, T. Lithgow, *Mol. Biol. Cell* **19**, 126 (2008).
8. J. G. Sklar, T. Wu, D. Kahne, T. J. Silhavy, *Genes Dev.* **21**, 2473 (2007).
9. S. Kutik *et al.*, *Cell* **132**, 1011 (2008).
10. K. Yamano, S. Tanaka-Yamano, T. Endo, *EMBO Rep.* **11**, 187 (2010).
11. J. G. Wideman *et al.*, *Mol. Biol. Cell* **10.1091/mbcE09-10-0844** (2010).
12. N. Thornton *et al.*, *J. Mol. Biol.* **396**, 540 (2010).
13. D. M. Walther, D. Papic, M. P. Bos, J. Tommassen, D. Rapoport, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 2531 (2009).
14. V. Robert *et al.*, *PLoS Biol.* **4**, e377 (2006).
15. B. Kornmann *et al.*, *Science* **325**, 477 (2009).

10.1126/science.1190507

ECOLOGY

Are Lizards Toast?

Raymond B. Huey,¹ Jonathan B. Losos,² Craig Moritz³

Lizards should be relatively invulnerable to warming: They are very good at evading thermal stress, tolerate high body temperatures, and resist water loss. Nevertheless, on page 894 of this issue, Sinervo *et al.* (1) document extinctions of lizard populations on five continents and argue that global warming is responsible. They use a simple biological model, validated against observed extinctions, to predict that warming will drive almost 40% of all global lizard populations extinct by 2080. If their prediction is even close to correct, lizards may be “the new amphibians” (2) in a race toward extinction.

A stark result for a genus of lizards in México leads off their paper: 12% of 200 previously validated *Sceloporus* populations (all with intact habitats) went extinct in recent

decades. Moreover, extinction probability was correlated with magnitude of warming at that site in spring, but not in other seasons. This correlation suggests that extinction is driven by energetic shortfalls during spring (when reproductive energy demands are highest), rather than by summer heat stress. Lizard natural history is instructive here: On hot days, lizards seek cooler refuges, such as burrows. With warming, lizards will spend longer periods in refuges, reducing foraging time, such that net energy gain becomes insufficient for reproduction; extinction ensues.

To test this mechanistic hypothesis, Sinervo *et al.* examined four *S. serrifer* populations, two of which have recently gone extinct (see the first figure). Using field estimates of maximum available body temperatures of lizards (operative temperatures) at these sites in spring and of body temperatures acceptable for activity, they predicted the number of hours per day that operative temperatures exceeded a lizard's thermal preferences, thus forcing retreat (see the second figure). At sites

Warming is held responsible for a rash of extinctions of global lizard populations.

where the lizards are now extinct, predicted time restrictions exceeded 3.85 hours; but at sites where lizards persist, predicted restrictions were shorter. Sinervo *et al.* then used air temperature data from weather stations to estimate time restrictions at all Mexican sites. *Sceloporus* populations with predicted restrictions above 3.85 hours in spring had higher extinction rates than did populations with shorter restrictions.

To predict future extinctions, Sinervo *et al.* applied their history-validated approach to current and future warming scenarios across the globe, using 1216 lizard populations on four continents. First, by resurveying known lizard populations and conducting literature surveys, they detected many extinctions; for example, 21% of Madagascar lizard populations in nature reserves have gone extinct. Estimated activity-time restrictions (with critical thresholds tuned to the thermal biology of each lizard family) effectively predicted populations that had gone extinct. Based on these data, the authors estimate that by 2080,

¹Department of Biology, University of Washington, Seattle, WA 98195, USA. ²Museum of Comparative Zoology, Harvard University, Cambridge, MA 02138, USA. ³Museum of Vertebrate Zoology, University of California, Berkeley, CA 94720, USA. E-mail: hueyrb@uw.edu



No escape from warming. The lizard *Sceloporus serrifer* has gone locally extinct at several warmed sites in México.

39% of all lizard populations and 20% of all lizard species will be extinct.

These observations and projections are disturbing, and concordance between predicted and observed extinctions on different continents suggests that they should be taken seriously. Other studies also warn that reptiles are vulnerable to warming (3–5). Even so, cautious skepticism is prudent, and some aspects of the study warrant further investigation.

How strong is the evidence of local extinctions? Lizard populations rise and fall over time, and failure to detect individuals during short surveys may indicate transient rarity rather than extinction. The conspicuousness of *Sceloporus* and other highlighted species argues against such “pseudo-extinction,” but only follow-up surveys can resolve whether these are true local extinctions.

How strong are the analytical approaches? The global scale of this study demanded methodological compromises, but more robust ecophysiological approaches are available (5–7). The prediction of widespread

that even a 2°C increase in air temperature will severely restrict activity time, reducing energy gain and rates of population growth, and thus precipitate extinction (6).

Might lizards be able to escape this toaster oven? Rapid genetic responses to climate warming have been documented in insects (9), but seem less likely in organisms like lizards with longer generation times (10). Using genetic models, Sinervo *et al.* conclude that genetic adaptation is not feasible. Some species may evade extinction by retreating uphill or to higher latitudes, where operative temperatures are lower (11, 12). However, human-induced habitat fragmentation may block such moves, and montane species may eventually run out of space (1). Moreover, observed extinctions of several montane *Sceloporus* populations appear related to increased interspecific competition from upward movements of lowland species (1).

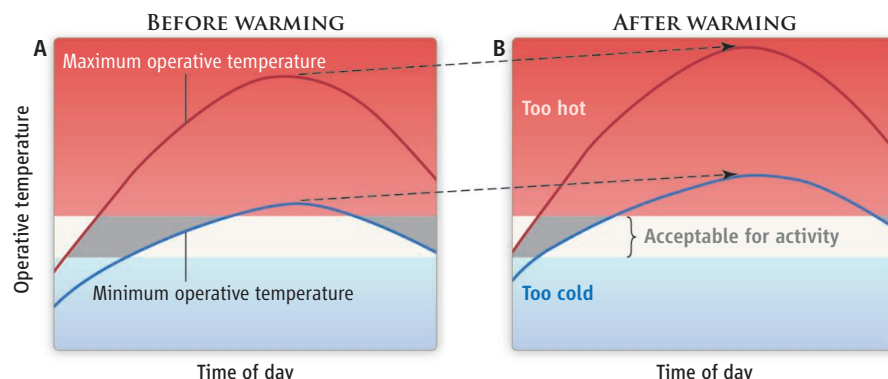
Lizards are remarkably diverse in geography, physiology, behavior, and habitat (13). Are Mexican *Sceloporus* reliable models for

lizards? We think that they are: Most lizards are also diurnal, thermoregulate carefully, and are active at high body temperatures, and most lineages show limited variation in thermal physiology between species. However, tropical forest lizards are different (4), because they do not thermoregulate carefully and are active at low body temperatures; and some lineages—most famously *Anolis* lizards (14)—show marked variation in thermal physiology between species. Whether they have the genetic capacity to outrun global warming (10) remains to be evaluated, but biophysical and physiological data suggest that even these species are at risk (4).

Global warming is expected to drive widespread extinctions, but predictions are rarely validated against actual extinctions and by knowledge of causal mechanisms (15). Sinervo *et al.* deliver a disturbing message: Climate-forced extinctions are not only in the future but are happening now. Moreover, the authors provide an effective framework for exploring organismal susceptibility to climate change. The steps involve documenting extinctions, evaluating underlying biophysical and eco-physiological mechanisms, considering the potential for adaptive evasion, and then building projection models based explicitly on established mechanisms. This should be the logical framework even as more complex and sophisticated methodologies are applied.

References

1. B. Sinervo *et al.*, *Science* **328**, 894 (2010).
2. D. B. Wake, V. T. Vredenburg, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 11466 (2008).
3. S. M. Whitfield *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 8352 (2007).
4. R. B. Huey *et al.*, *Proc. R. Soc. Lond. B. Biol. Sci.* **276**, 1939 (2009).
5. N. J. Mitchell, M. R. Kearney, N. J. Nelson, W. P. Porter, *Proc. R. Soc. Lond. B. Biol. Sci.* **275**, 2185 (2008).
6. A. E. Dunham, in *Biotic Interactions and Global Change*, P. M. Kareiva, J. G. Kingsolver, R. B. Huey, Eds. (Sinauer, Sunderland, MA, 1993), pp. 95–119.
7. L. B. Buckley, *Am. Nat.* **171**, E1 (2008).
8. S. C. Adolph, W. P. Porter, *Am. Nat.* **142**, 273 (1993).
9. W. E. Bradshaw, C. M. Holzapfel, *Science* **312**, 1477 (2006).
10. L.-M. Chevin, R. Lande, G. M. Mace, *PLoS Biol.* **8**, e1000357 (2010).
11. R. K. Colwell, G. Brehm, C. L. Cardelús, A. C. Gilman, J. T. Longino, *Science* **322**, 258 (2008).
12. C. Moritz *et al.*, *Science* **322**, 261 (2008).
13. E. R. Pianka, L. J. Vitt, *Lizards: Windows to the Evolution of Diversity* (Univ. of California Press, Berkeley, 2003).
14. J. B. Losos, *Lizards in an Evolutionary Tree: Ecology and Adaptive Radiation of Anoles* (Univ. of California Press, Berkeley, CA, 2009).
15. C. Parmesan, *Annu. Rev. Ecol. Syst.* **37**, 637 (2006).



Time is of the essence. Warming will shorten activity times of lizards, potentially reducing energy gains below levels required for reproduction and thus causing extinction. (A) Maximum (red lines) and minimum (blue lines) operative temperatures of lizards during a spring day before global warming. Lizards are active whenever the operative temperatures are within an acceptable range (gray fill). (B) Operative temperatures rise after warming, shortening lizard activity time.

PHOTO CREDIT: FAUSTO MÉNDEZ-DE-LA-CRUZ

MOLECULAR BIOLOGY

Small RNA Makes Its Move

Rob Martienssen

It has been known for almost 100 years that when a plant virus infects a leaf, mobile signals are transmitted through vessels in the stem to other leaves to confer resistance to subsequent infection. More recently, the silencing of exogenous transgenes has been shown to involve a mobile signal (1). Although RNA molecules have been implicated in systemic plant cell-to-cell communication, the nature of mobile RNA that silences gene expression has not been clear (2). Now, four studies—including those by Molnar *et al.* (3) and Dunoyer *et al.* (4) on pages 872 and 912 of this issue—report that small interfering RNA (siRNA) and microRNA (miRNA) are mobile signals that control gene expression during plant development.

Two of the studies examined the genetic pathway required for the mobile silencing signal by making grafts between wild-type and mutant roots and shoots of the model plant *Arabidopsis thaliana*. Molnar *et al.* used next-generation sequencing to detect mobile small RNA in grafted roots defective in DCL2, DCL3, and DCL4, the Dicer-like enzymes in *Arabidopsis* that cleave long precursor RNA into small RNA. Thus, these roots cannot generate 22- and 24-nucleotide (nt) siRNA. Nonetheless, the roots accumulated these forms, indicating siRNA movement from the shoot. Small RNA generated from hundreds of loci was found in the grafted roots, mostly from transposons (DNA sequences that can move around the genome). By contrast, in another study, Dunoyer *et al.* (5) identified small RNA in grafted roots that derived from inverted repeats in the shoot.

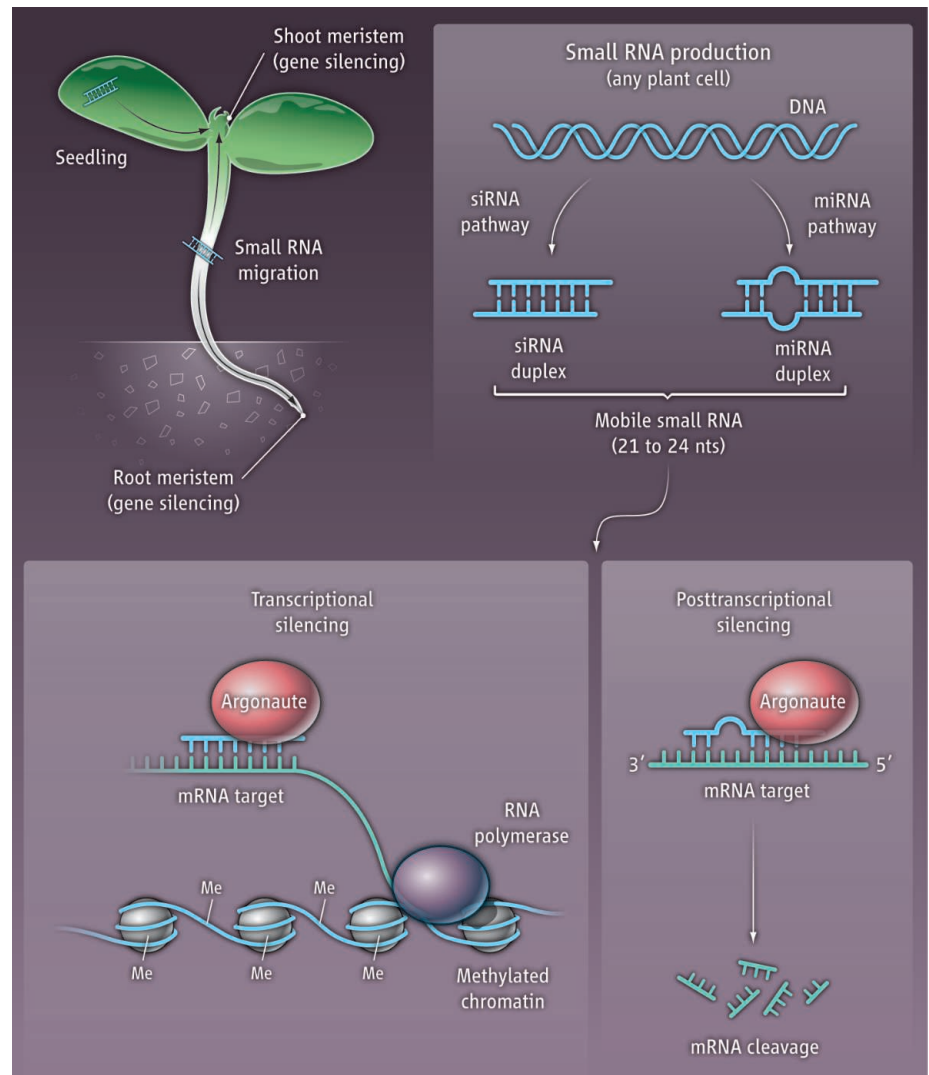
Dunoyer *et al.* (4) further show that 21-nt siRNA is mobile by introducing small RNA directly (by particle bombardment) into *Arabidopsis* cells lacking RNA-dependent RNA polymerase. They observed the spread of targeted gene silencing along with the movement of labeled small RNA. The authors demonstrate that it is the double-stranded duplex, rather than the single-stranded form, of siRNA that is mobile. Dunoyer *et al.* (5) also report similar results with 24-nt siRNA duplexes.

All 3 classes of siRNA (21, 22, and 24 nt) are mobile and use components of both the

posttranscriptional and the transcriptional gene silencing pathways of RNA interference (RNAi) to silence their targets (1). These components include the plant-specific RNA polymerase Pol IV, two RNA-dependent RNA polymerases (RDR2 and RDR6), and the chromatin-remodeling protein CLASSY (1). The key nucleases required for RNAi—Dicer and Argonaute—are encoded by multigene families, and act redundantly to some extent in generating these mobile small RNA types. Dunoyer *et al.* (5) further dem-

onstrate that 24-nt mobile siRNA generated from inverted repeats depends on the exportin protein HASTY for their transport out of the nucleus. Importantly, Molnar *et al.* show that the targets of mobile siRNA signals are methylated by RNAi-dependent DNA methylation. This suggests that mobile RNA can have long-lasting effects on DNA silencing.

The rules for selecting which loci in the *Arabidopsis* genome generate mobile siRNA are still unclear. Naturally occurring inverted repeats are a major source of



Move to silence. A model for silencing gene expression during plant development involves mobile small RNA molecules (siRNA) that move systemically through the plant vascular system, as well as miRNA that may move over shorter distances. Small RNA move from source cells (such as in the shoot) into dividing cells in the root and shoot meristems where they silence targets, including transposons and genes.

Cold Spring Harbor Laboratory, Cold Spring Harbor NY 11724, USA. E-mail: martienss@cshl.edu

CREDIT: C. BICKEL/SCIENCE

mobile siRNA (5), and some of these have known silencing functions, such as the phosphoribosylanthranilate isomerase (PAI1-4) inverted repeat (3), which silences a functional copy of the same gene (6). These inverted repeats are polymorphic, indicating that they arise and disappear rapidly during evolution, presumably responding to selection for target gene expression. In maize, a naturally occurring, polymorphic inverted repeat comprises the promoter and part of the coding region of the *Mutator* (*Mu*) transposon gene and is expressed in the shoot tip. It generates 24- and 25-nt small RNA, which are found in the surrounding leaves, suggesting their mobility (7). Once silenced by these small RNA, *Mu* elements remain silent in subsequent generations in the presence of RDR2 and Pol IV.

Endogenous mobile small RNA may be generated by many potential cellular sources in plants, and act as signals during development of various target cells. For example, 21-nt trans-acting siRNA (tasiRNA) can move across young leaves to reach their targets, which are silenced posttranscriptionally (1, 4). Carlsbecker *et al.* now report that some 21-nt miRNA generated in root cells are mobile within the root, and posttranscriptionally silence multiple transcription factor genes in adjacent cells (8). These same miRNA appear to move in the shoot in the absence of the nuclease Argonaute1 (9), consistent with a mobile duplex form of small RNA.

Analogous to the transport of photosynthetic sugars, roots may act as a “sink” for mobile small RNA that originate in a “source” tissue, such as leaves of the shoot (3, 5). However, only lateral roots are affected, which emerge after silencing is initiated by shoot grafting (3). Similarly, when shoots are the recipient of silencing signals, only leaves that arise after root grafting are affected (2). This indicates that gene silencing occurs predominantly in undifferentiated stem cells (dividing cells in the shoot and root meristems) that give rise to new organs in plants (2). Perhaps this is because RNAi-mediated transcriptional silencing is most efficient in dividing cells (10).

Mobile 24-nt siRNA may account for some curious properties of transposon silencing during plant development. For example, silencing of *Mu* transposons in maize occurs in the shoot meristem, giving rise to clonal regions in leaves. As the plant matures, the number of sectors in each leaf increases (11), suggesting that the developmental transition between early and late leaves may be controlled by a silencing process, and that a mobile signal silences stem cells as the mer-

istem ages (12). These cells give rise to flowers, and so silencing is inherited in the next generation (11).

Flowers are also well-known sink tissues. Translocation of small RNA into flowers could affect the inheritance of epigenetic alleles. Sperm cells are loaded with mobile 21-nt transposon-targeting siRNA from the surrounding pollen grain, whereas the ovule and embryo sac have predominantly maternal 24-nt siRNA, which are required to silence transposons (13, 14) and inhibit germ cell fate in adjacent cells (13). Those small RNA derived from the plant body could find their way into ovules and pollen grains, which are physiological sinks, just like meristems and roots.

Small RNA silencing in eukaryotes is uniquely sensitive to temperature and potentially other environmental signals (10), providing a potential mechanism for the environment to modify the germ line. The inheritance of acquired characters was envisioned not only by Lamarck, but also by Darwin, who proposed the existence of gemmules,

somatic particles that entered the germ line and contributed such characters to the next generation (15). It remains to be seen if mobile small RNA will live up to that extraordinary vision.

References

1. P. Dunoyer, O. Voinnet, *Trends Plant Sci.* **14**, 643 (2009).
2. C. A. Brosnan *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 14741 (2007).
3. A. Molnar *et al.*, *Science* **328**, 872 (2010); published online 22 April 2010 (10.1126/science.1187959).
4. P. Dunoyer *et al.*, *Science* **328**, 912 (2010); published online 22 April 2010 (10.1126/science.1185880).
5. P. Dunoyer *et al.*, *EMBO J.* 10.1038/emboj.2010.65 (2010).
6. J. Bender, G. R. Fink, *Cell* **83**, 725 (1995).
7. R. K. Slotkin, M. Freeling, D. Lisch, *Nat. Genet.* **37**, 641 (2005).
8. A. Carlsbecker *et al.*, *Nature* 10.1038/nature08977 (2010).
9. C. A. Kidner, R. A. Martienssen, *Nature* **428**, 81 (2004).
10. A. Kloc, R. Martienssen, *Trends Genet.* **24**, 511 (2008).
11. R. Martienssen, A. Baron, *Genetics* **136**, 1157 (1994).
12. R. S. Poethig, *Science* **250**, 923 (1990).
13. V. Olmedo-Monfil *et al.*, *Nature* **464**, 628 (2010).
14. R. K. Slotkin *et al.*, *Cell* **136**, 461 (2009).
15. C. Darwin, *The Variation of Animals and Plants Under Domestication* (John Murray, London, 1868).

10.1126/science.1190510

PHYSICS

Managing Multistate Quantum Entanglement

Christoph Wildfeuer

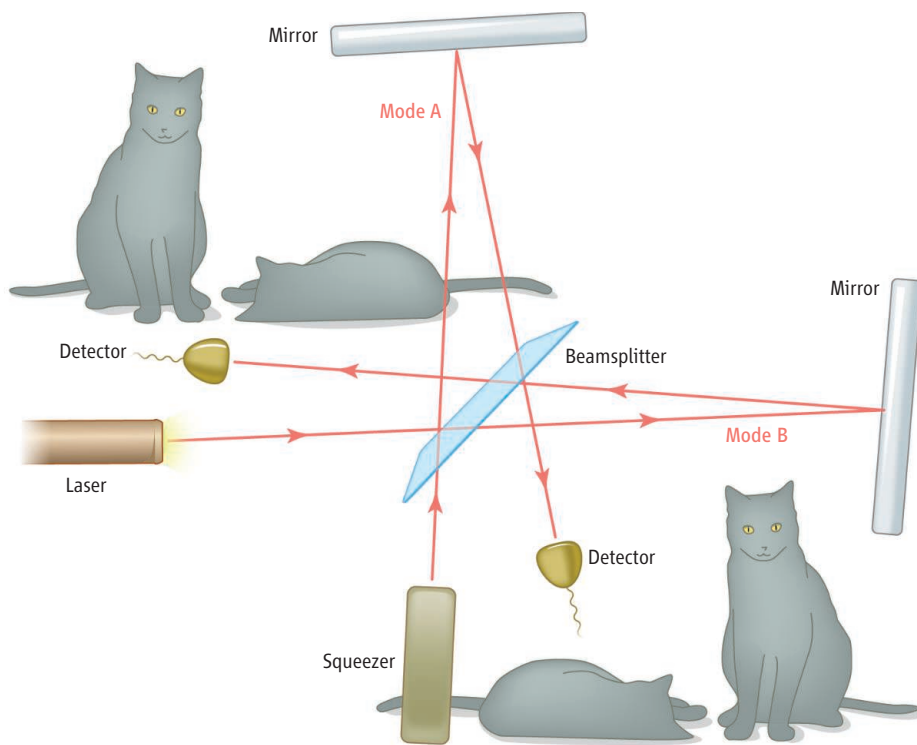
A general route for creating entangled states of large numbers of photons may be relevant to other applications of interferometry, such as gravity wave detection.

Good experiments minimally perturb the sample under study, but in the quantum realm, measurement cannot be separated from the system. Before the measurement, all possible outcomes can form an entangled state; measurement then causes this entangled state to collapse to one outcome. In 1935, Schrödinger presented a thought experiment to show how strange this situation really is. A cat is hidden in a box along with a sample of radioactive nuclei. If a nuclear decay occurs, poison is released and the poor cat dies. However, until we look in the box, the state of the cat entangles both a live and a dead cat. Experiments with Schrödinger-cat states that entangle several particles continue to provide deep insights into the measurement process. One

such state, called a NOON state, entangles a fixed number of photons N , all of which are in one of two possible states (if these were coins, they would be all heads or all tails). On page 879 of this issue, Afek *et al.* (1) achieved a record by making NOON states in an interferometer with five entangled photons. Their approach may have implications for other implementations of interferometry, such as gravity wave detectors.

The NOON state was first discussed in 1989 by Sanders, who was particularly interested in the Schrödinger-cat aspect and quantum decoherence—how entangled states decay over time (2). These states can be the different optical paths, or modes, in an interferometer (see the figure). In the NOON state, all of the photons are mode a or mode b . However, we cannot tell which mode is occupied. In 2000, Dowling's group rediscovered this state in the context of quantum imaging, particularly for using quantum interference to improve the res-

Department of Physics and Astronomy, Louisiana State University, Baton Rouge, LA 70803, USA. E-mail: wildfeuer@phys.lsu.edu



Arms that cradle Schrödinger's cat. An important goal in quantum mechanics is to create states that entangle many particles. The NOON state puts several photons all in one of two possible states (in this case, into one of two arms of an interferometer: Afek *et al.* actually use a Mach-Zehnder interferometer, which is mathematically equivalent to the Michelson interferometer displayed here). The states are entangled, so until a measurement is performed, we do not know which arm has the photons (the two modes, *a* and *b*, correspond to the two states, alive or dead, of Schrödinger's cat). Afek *et al.* delivered laser light photons into one mode of the first beamsplitter, and highly correlated (or squeezed) photons into the other. Their interference created a state that is very close to being a pure NOON state for five photons.

olution limits in lithography (3). A few experiments have reported $N = 4$ NOON states (4). Afek *et al.* have now pushed the limit to $N = 5$, but what is even more remarkable is that their generation scheme should be scalable to any photon number N .

Making a NOON state with 100% fidelity is quite a cumbersome problem for $N \geq 3$, as it requires complicated multistep protocols that usually have poor efficiencies. The road to success for Afek *et al.* was to use an experimental approach that relaxed the requirement to make the NOON state exactly, yet retained all of its useful properties. The theoretical underpinnings of the approach had been published by Hofmann and Ono in 2007 (5). Statistical uncertainty in the number of photons in a laser pulse causes noise in the signal of an interferometer. A quantum mechanical trick called squeezing reduces this uncertainty; hence, there is less noise in an interferometer with squeezed light. The photon number distribution actually becomes narrower than Poissonian (i.e., sub-Poissonian).

Squeezed photons can be produced in an optical crystal by a technique called paramet-

ric down-conversion. For instance, a blue photon is converted into two red photons, with each red photon having just half the energy of the original blue photon. The down-converted photons appear in pairs and their behavior is more highly correlated than that of the initial blue photons. In Afek *et al.*'s approach, the coherent photons from a weak laser pulse and the squeezed photon pairs combine at a beamsplitter, where they can undergo constructive or destructive interference.

This interference results in a modulation of the two-photon (or, in general, multiphoton) distribution. Constructive interference will lead to photon bunching (the sub-Poissonian distribution mentioned above), whereas destructive interference will lead to photon anti-bunching (wider than Poissonian, or super-Poissonian). Hence, we get a better (sub-Poissonian) measurement, but only part of the time, so the Heisenberg uncertainty principle still holds (6). A rigorous calculation based on the experimental results shows that the superpositions of multi-photon number states are very close to an ideal NOON state. Surprisingly, the

scheme presented by Afek *et al.* is similar to an idea of Caves in 1981 to improve the sensitivity of gravitational wave detectors (7). He showed that by squeezing the zero-point energy state of an interferometer, the noise of the circulating laser beam can be reduced. The sensitivity of the interferometer is improved below its classical limit, called the shot-noise limit. This technique may soon be applied at the Laser Interferometer Gravitational Wave Observatory (LIGO), which consists of two large L-shaped Michelson interferometers (8).

What Afek *et al.* have shown is that squeezing photons at LIGO may generate large NOON (or "high-NOON") states in the arms of the interferometer. Hence, the photons become path-entangled, which creates stronger-than-classical intensity correlations between the two arms of the interferometer. This process improves the sensitivity beyond the classical shot-noise limit. This limit can also be decreased by increasing the optical power inside the interferometer, but this option can cause thermal problems. LIGO now uses ordinary photodiodes for detecting the signal of the interferometer, and exceeding the shot-noise limit with Afek *et al.*'s approach requires incredibly sensitive detectors that can resolve changes of just one photon. However, scientists at the National Institute of Standards and Technology have developed such a detector that can resolve up to 20 photons with a detection efficiency of more than 95% (9).

Afek *et al.*'s approach may already prove advantageous for interferometry below the shot-noise limit, where the power in the interferometer is limited by material constraints. The applications of this method for improving the sensitivity of gravity wave detection may still be far off in the future. Nonetheless, this idea represents a wonderful example in which quantum mechanics provides an understanding of the predictions of a classical field theory, namely Einstein's theory of general relativity.

References

1. I. Afek, O. Ambar, Y. Silberberg, *Science* **328**, 879 (2010).
2. B. C. Sanders, *Phys. Rev. A* **40**, 2417 (1989).
3. A. N. Boto *et al.*, *Phys. Rev. Lett.* **85**, 2733 (2000).
4. J. P. Dowling, *Contemp. Phys.* **49**, 125 (2008).
5. H. F. Hofmann, T. Ono, *Phys. Rev. A* **76**, 031806(R) (2007).
6. Y. J. Lu, Z. Y. Ou, *Phys. Rev. Lett.* **88**, 023601 (2001).
7. C. M. Caves, *Phys. Rev. D* **23**, 1693 (1981).
8. H. J. Kimble, Y. Levin, A. B. Matsko, K. S. Thorne, S. P. Vyatchanin, *Phys. Rev. D* **65**, 022002 (2002).
9. A. E. Lita, A. J. Miller, S. W. Nam, *Opt. Express* **16**, 3032 (2008).

10.1126/science.1190491

MOLECULAR BIOLOGY

Phylogeny of Methylomes

Albert Jeltsch

The DNA of most species is methylated, containing the modified base 5-methylcytosine. This modification has a role in silencing gene expression, among other important functions (1, 2). Advances in sequencing methods have allowed measurement of the first complete genome-wide DNA methylation map ("methylome") of the model plant *Arabidopsis thaliana* and human cells (3–5). Studies by Feng *et al.* (6) and by Zemach *et al.* on page 916 of this issue (7) now expand this list by providing genome-wide methylomes for 20 additional species, revealing important conserved features and phylogenetic relationships of the methylation machinery.

Human DNA shows a dense and genome-wide methylation at CpG sites (cytosine and guanine dinucleotides) (1–3), including methylation of transposable elements, which suppress their transcription and recombination. Unmethylated CpG sites are mainly located in so-called CpG islands, regions with large CpG content. These islands often lie in the promoter regions of genes, and if methylated, they repress gene expression. Plants share features with mammalian methylation, such as methylation of transposons and lack of methylation at promoters of expressed genes, but they possess three methylation systems specific for CpG, CHG, and CHH (where H is the nucleotide adenine, thymine, or cytosine). In addition, methylation of gene bodies (the exons and introns of the genes) has been observed in mammals and plants (3, 4, 8, 9). However, not all species show high DNA methylation: The eukaryotic model organisms *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster*, and *Caenorhabditis elegans* show zero or very little DNA methylation. As a consequence, DNA methylation was long considered an interesting phenomenon but not of general importance.

For animals, Feng *et al.* and Zemach *et al.* show strong preference for methylation at CpG sites due to high specificity of the DNA methyltransferase 1 (Dnmt1) enzymes. Non-CpG methylation is introduced by Dnmt3 enzymes. Fungi show strong non-CpG methylation by a subclass of Dnmt1 enzymes, fungi-specific

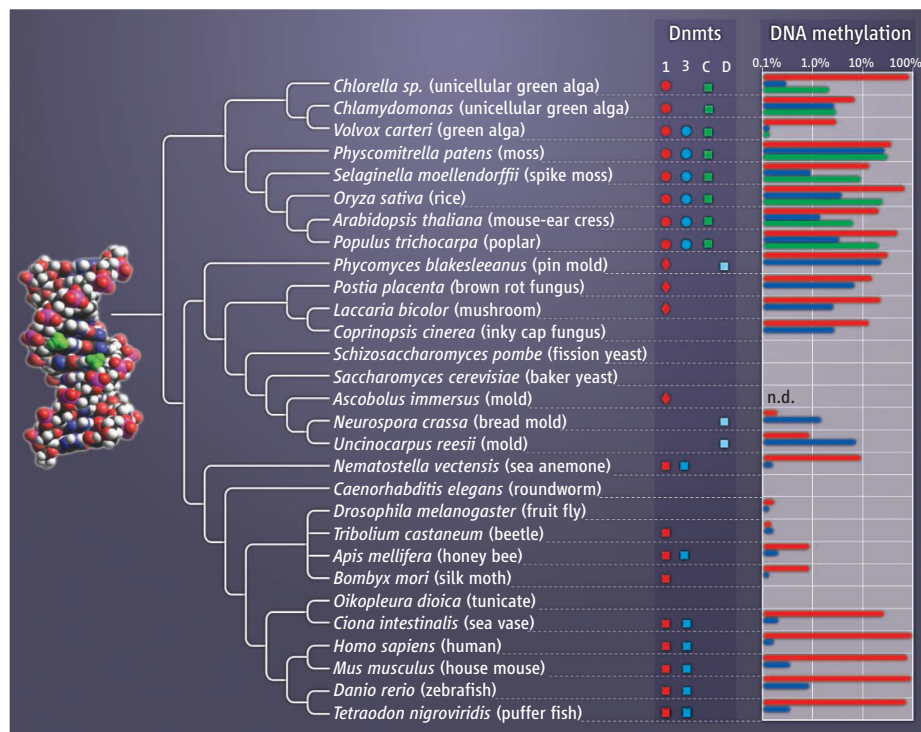
Dim-2-type enzymes, and perhaps additional families of Dnmts. In plants, CpG methylation is maintained by the plant subfamily of Dnmt1 enzymes called Met1; CHG methylation is mediated by the plant-specific chromomethylases; and CHH methylation is introduced by the DRM enzymes (see the figure) the plant subfamily of Dnmt3 enzymes. Conservation of the different Dnmts and DNA methylation levels suggests that the last common ancestor of eukaryotes contained a functional DNA methylation system with secondary expansion and the loss of methylation in some lineages. Primitive methylation likely occurred at low to intermediate levels and was targeted to gene bodies and transposable elements, leaving gene promoters unmethylated.

Zemach *et al.* and Feng *et al.* found that methylation of the gene body (3, 4, 8, 9) is highly conserved and likely was present in the last common ancestor. It might be involved in the prevention of transcriptional initiation within the gene body. Interestingly, gene body

The patterns of DNA methylation across the genomes of 20 eukaryotes reveal conserved features and specific roles during evolution.

methylation shows a parabolic correlation with gene expression, with genes expressed at an intermediate level having the highest amounts of gene body methylation (7). This suggests the coexistence of two opposing targeting mechanisms that increase or decrease methylation with gene expression. Two interesting details of this process are that exons tend to be more highly methylated than introns (as in human DNA) (3, 10) and that the end of the gene shows a similar drop in methylation as the gene's promoter region. These findings imply general roles of DNA methylation in transcriptional elongation, termination, and perhaps alternative splicing.

One trend observed in all kingdoms is that highly developed multicellular organisms show increased DNA methylation. This may be due to the need for more efficient control of transposons because of outbreeding sexual propagation (7), or the need for additional epigenetic regulation to control the development of many different cell types. Lower methyla-



Methylomes and methylation machineries. The phylogenetic tree was made with National Center for Biotechnology Information taxonomy and the Interactive Tree of Life. Dnmt1 (red), Dnmt3 (blue), chromomethylase (green), and Dim-2 families (light blue) are shown. DNA methylation data (6, 7) are averaged. Only plant CMT methyltransferases display a preference for CHG sites; CHG methylation is shown only for plants. Methylation data for human (10) and *N. crassa* (23) are included. *A. immersus* contains DNA methylation (24), but genome-wide distribution has not been determined (n.d.). DNA methylated at a CpG site (green) is on the left.

tion of gene promoter regions is observed in most species. It is correlated to gene expression in plants and animals, which suggests that the role of promoter methylation in gene regulation may be derived. In mammals and flowering plants, this process has been developed into the imprinting system that generates parent of origin-dependent methylation marks on a small subset of genes and leads to their monoallelic expression (11, 12).

On the other hand, several species underwent loss of DNA methylation to a large degree, such as the nematode *C. elegans*, the insect *D. melanogaster*, and the yeasts *S. pombe* and *S. cerevisiae*; hence, DNA methylation is not essential under all conditions of life, and its loss may even be beneficial. This could be connected to the resources allocated to global DNA methylation or to the fact that 5-methylcytosine is mutagenic, because its deamination is repaired less efficiently than deamination of unmethylated cytosine.

Human DNA methylation patterns vary with cell type and developmental stage (3, 10, 13–18), with disease (10, 15, 16), between alleles (11, 19–22), and among individuals (14, 15). Future methylome studies may reveal similar features in the other model organisms as well. Functional investigations of the properties of Dnmts and their interaction with chromatin and additional factors will clarify the mechanisms by which methylation patterns are set and maintained. Other questions yet to be resolved surround the means by which repeats and transposable elements are identified, the processes that target DNA methylation to gene bodies in an expression-dependent manner, the biological function of gene body methylation that led to its high stability in evolution, and the mechanisms to generate and modify tissue-specific patterns of promoter methylation in mammals.

References

1. M. M. Suzuki, A. Bird, *Nat. Rev. Genet.* **9**, 465 (2008).
2. J. A. Law, S. E. Jacobsen, *Nat. Rev. Genet.* **11**, 204 (2010).
3. R. Lister *et al.*, *Nature* **462**, 315 (2009).
4. S. J. Cokus *et al.*, *Nature* **452**, 215 (2008).
5. R. Lister *et al.*, *Cell* **133**, 523 (2008).
6. S. Feng *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 101073/pnas.1002720107 (2010).
7. A. Zemach *et al.*, *Science* **328**, 916 (2010); published online 15 April 2010 (10.1126/science.1186366).
8. M. P. Ball *et al.*, *Nat. Biotechnol.* **27**, 361 (2009).
9. X. Zhang, *Science* **320**, 489 (2008).
10. E. Hodges *et al.*, *Genome Res.* **19**, 1593 (2009).
11. W. Reik, J. Walter, *Nat. Genet.* **27**, 255 (2001).
12. P. E. Jullien, F. Berger, *Curr. Opin. Plant Biol.* **12**, 637 (2009).
13. A. Meissner *et al.*, *Nature* **454**, 766 (2008).
14. Y. Zhang *et al.*, *PLoS Genet.* **5**, e1000438 (2009).
15. S. Maegawa *et al.*, *Genome Res.* **20**, 332 (2010).
16. R. A. Irizarry *et al.*, *Nat. Genet.* **41**, 178 (2009).
17. R. Straussman *et al.*, *Nat. Struct. Mol. Biol.* **16**, 564 (2009).
18. C. Popp *et al.*, *Nature* **463**, 1101 (2010).
19. K. Kerker *et al.*, *Nat. Genet.* **40**, 904 (2008).
20. B. Panning, *J. Biol.* **7**, 30 (2008).
21. Y. Zhang *et al.*, *Genome Biol.* **10**, R138 (2009).
22. L. C. Schalkwyk *et al.*, *Am. J. Hum. Genet.* **86**, 196 (2010).
23. K. A. Borkovich *et al.*, *Microbiol. Mol. Biol. Rev.* **68**, 1 (2004).
24. C. Goyon, T. I. Nogueira, G. Faugeron, *J. Mol. Biol.* **240**, 42 (1994).

10.1126/science.1190738

CLIMATE

A Paleoclimatic Enigma?

William F. Ruddiman

Major glaciations began in the Northern Hemisphere around 2.75 million years ago, after a long prior interval of climatic cooling. Numerous observations reveal how climate cooled before glacial onset, but our understanding of the driving forces behind the cooling remains incomplete.

Climate had been cooling from pole to pole for 50 million years before northern glacial onset (see the figure). Arctic forests changed from frost-intolerant evergreens to temperate deciduous trees to cold-adapted spruce and larch and eventually to tundra near the time of glacial onset (1). Antarctica was mostly ice-free until 34 million years ago; glaciers of varying size then existed on the continent until 14 million years ago, after which a large and relatively stable ice sheet formed. The gradual shift toward heavier $\delta^{18}\text{O}$ values in CaCO_3 shells of sea-floor foraminifera since 50 million years ago documents a combined deep-ocean cooling and increase in Antarctic ice (2, 3).

Until a decade ago, most paleoclimate modelers attributed this ongoing bipolar cooling to a gradual reduction in the CO_2 con-

centration in the atmosphere. This inferred CO_2 decrease was ascribed to a combination of reduced volcanic CO_2 input to the ocean and atmosphere because of a slowing rate of sea-floor spreading (4) and increased CO_2 removal by enhanced chemical weathering in tectonically uplifting regions like Tibet (5).

Methods devised to reconstruct the CO_2 concentration of the atmosphere on longer time scales later tested this conclusion. One method relied on the ^{13}C composition of marine organic molecules called alkenones (6), another on boron-isotopic values in marine carbonate sediments (7). Results from both methods suggested estimated CO_2 concentrations of around 1000 parts per million (ppm) or more during much warmer climates tens of millions of years ago, compared with ice-core values of just 180 to 300 ppm during the glacial cycles of the past 800,000 years.

In a broad sense, this long-term CO_2 decrease provided some support for the idea that CO_2 has been the long-term driver of global cooling, but a closer look revealed major problems. By 22 million years ago, the alkenone and boron isotope data both showed that estimated CO_2 concentrations were already within the range typical of the glacial cycles of the past 800,000 years. If CO_2 concentrations of 180 to 300 ppm have played an integral role in allowing glacial cycles in the

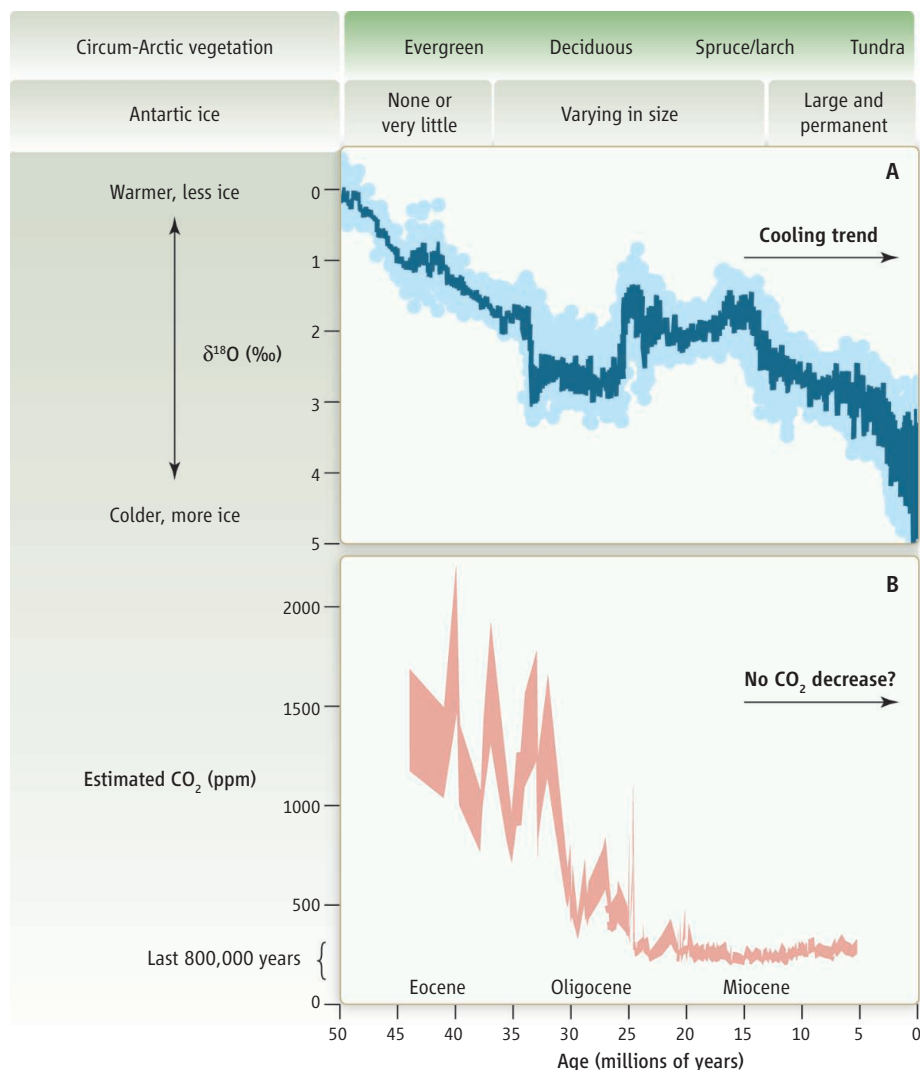
What caused the onset of major Northern Hemisphere glaciation about 2.75 million years ago?

past 800,000 years, why did comparably low CO_2 values 22 million years ago not initiate glacial cycles? And if the average CO_2 trend has not fallen in the past 22 million years, what caused the substantial bipolar cooling during that time?

Other proposed causes seem insufficient to explain large-scale cooling. Gradual plate motions and falling sea level have extended the northern margins of circum-Arctic continents into cooler near-polar latitudes (8), but models suggested that these factors were not enough to explain the major cooling observed. Closing of the Isthmus of Panama, about two million years before the onset of northern glaciations, has been proposed as a causal factor in glacial inception. Results from coupled ocean-atmosphere models indicate, however, that isthmus closure would have sent greater amounts of sensible heat northward in the Atlantic and melted more snow and ice, rather than promoting ice growth by delivering more moisture (9).

In the past 15 million years, broad-scale uplift of plateaus and mountains has occurred in the northern and eastern Tibetan Plateau, the east-central Andes and Altiplano, the east African rift valley, and the northern Canadian Rockies. Elevation of these rock surfaces to cooler levels in the atmosphere would have cooled the uplifted terrain and nearby regions,

Department of Environmental Sciences, University of Virginia, Charlottesville, VA 22903, USA. E-mail: wfr5c@virginia.edu



Indices of climate change in the past 50 million years. (A) Changes in circum-Arctic vegetation (1). (B) Changes in Antarctic ice sheet size. (C) Oxygen isotopic ($\delta^{18}\text{O}$) index of deep-ocean temperature and ice volume (2, 3). (D) Estimates of past CO_2 concentrations from alkenones (6). Paleoclimatic data from both poles indicate a progressive cooling from 50 million years ago until the onset of Northern Hemisphere glaciation 2.75 million years ago, but reconstructions of atmospheric CO_2 concentrations from two prominent proxy methods show very low values by 25 million years ago and no subsequent decrease. If the CO_2 trend did not fall after 25 million years ago, the cooling trend would be difficult to explain.

but apparently without causing a large cooling in far-off polar regions (10). Chemical weathering of eroded debris in these uplifted regions is a plausible means of reducing atmospheric CO_2 concentrations (5), but this explanation brings us back to the problem of the persistently low CO_2 concentrations estimated for the past 22 million years (see the figure).

Paleoclimatologists were left with three possibilities. First, we might have overlooked something crucial. This option seems unlikely for a science that has been growing toward maturity for decades.

Second, one or more of the physical mechanisms previously proposed to account for global cooling could have had a much stronger effect than thought. Slow north-

ward plate motions (8) and gradual plateau and mountain uplift (10) seem like good candidates to drive a gradual cooling, and in recent years, many interactive feedback processes added to general circulation climate models have boosted the size of climatic responses to these imposed forcings. Still, these feedbacks seem unlikely to have provided the amplification required.

Third, the proxy methods used to reconstruct CO_2 concentrations prior to ice-core records could be invalid. CO_2 concentrations estimated from the two proxies disagree by large amounts between 45 and 25 million years ago, with boron-isotope estimates having fallen to very low values but alkenone estimates still at high concentrations. Another

problem is a disagreement with climate model simulations, which indicate that CO_2 concentrations below ~ 750 ppm should cause a large ice sheet to form on Antarctica (11). By this measure, the CO_2 estimates from boron isotope data predict the appearance of a large ice sheet by 55 to 50 million years ago, some 20 million years before the onset of substantial Antarctic ice (see the figure). The CO_2 trend from the alkenone proxy predicts that a large ice sheet would have formed by 30 million years ago, a few million years after the time that ice masses of varying sizes first formed, but long before the appearance of a large and stable ice mass.

Recent data have now raised further questions about the older CO_2 proxy estimates. The new boron/calcium (B/Ca) ratio predicts CO_2 concentrations of 350 to 450 ppm between 20 and 10 million years ago, a time when the other two techniques indicated values of 200 to 300 ppm (12). It also predicts somewhat higher concentrations between 1.4 and 0.9 million years ago than recent results from the boron isotope method (13). And a recent alkenone-based study shows a substantial CO_2 decrease from 5 to 2 million years ago (14). In general, these newer results suggest that a gradual CO_2 decline did play a role in northern glacial inception, contrary to previous proxy estimates.

From a paleoclimatic perspective, the simplest way to explain the large and ongoing bipolar cooling during the past 50 million years remains what it was a decade or more ago: a progressive decrease in atmospheric CO_2 levels that began at high levels 50 million years ago and continued to fall until the past few million years, rather than "bottoming out" as a forcing factor about 22 million years ago. If this view is correct, geochemists still have work to do in refining the CO_2 proxies.

References

1. J. A. Wolfe, *Am. Sci.* **66**, 994 (1978).
2. K. G. Miller, R. G. Fairbanks, G. S. *Paleoceanography* **2**, 1 (1987).
3. J. Zachos et al., *Science* **292**, 686 (2001).
4. R. A. Berner et al., *Am. J. Sci.* **283**, 641 (1983).
5. M. E. Raymo et al., *Geology* **16**, 649 (1988).
6. M. Pagani et al., *Science* **309**, 600 (2005).
7. P. N. Pearson, M. R. Palmer, *Nature* **406**, 695 (2000).
8. E. J. Barron, W. M. Washington, in *The Carbon Cycle and Atmospheric CO_2 : Natural Variations, Archaeal to Present*, E. T. Sundquist, W. S. Broecker, Eds. (Geophys. Monogr. **32**, American Geophysical Union, Washington, DC, 1985), pp. 504–529.
9. U. Mikolajewicz et al., *Paleoceanography* **8**, 409 (1993).
10. W. F. Ruddiman, J. E. Kutzbach, *Sci. Am.* **264**, 66 (1991).
11. R. M. DeConto, D. Pollard, *Nature* **421**, 245 (2003).
12. A. K. Tripathi et al., *Science* **326**, 1394 (2009).
13. B. Hönisch et al., *Science* **324**, 1551 (2009).
14. M. Pagani et al., *Nat. Geosci.* **3**, 27 (2010).

10.1126/science.1188292



INTRODUCTION

Landscapes of Infection

Infectious disease remains one of the biggest killers in developing countries. Two of them account for an enormous toll: Eleven million people live with tuberculosis (TB), and almost 250 million cases of malaria—and roughly a million deaths among children—were reported in 2008; a staggering assault on humankind. The 33 million people who live with HIV/AIDS are frequently co-infected with TB and/or malaria, and co-infection increases the overall risk of mortality and morbidity from all three diseases. Once, we aspired to find “magic bullet” solutions to these plagues using vaccines or drugs, but we have learned that there are no cure-all or simple solutions. These pathogens have complex repertoires of genetic resources that permit them to constantly reinvent themselves and escape the pressures applied by infection-control measures. To curb these elusive targets, we, too, need a large repertoire of tools.

In this special issue of *Science*, two pairs of Reviews investigate the variety of cell, molecular, and epidemiological research strategies currently being used to understand and control malaria and TB. Kappe *et al.* (p. 862) describe aspects of the basic biology of malaria infection and how understanding it might someday contribute to the ambitious goal of eradication. Mackinnon and Marsh (p. 866) focus on the constantly shifting selective pressures acting on the most lethal form of the malaria parasite, *Plasmodium falciparum*. Russell *et al.* (p. 852) explore the basic biology of *Mycobacterium tuberculosis*, the infectious agent responsible for TB, and highlight the gaps in knowledge that hamper the ability to successfully treat the disease. Dye and Williams (p. 856) explore the population dynamics of TB at a time when, despite much success in providing effective therapy, disease burden remains enormous and highly drug-resistant strains are emerging.

The News section focuses exclusively on malaria, chronicling the remarkable successes in malaria control in Africa in the past few years and the huge remaining challenges (p. 842). These gains are in jeopardy, however, from resistance to artemisinin-based combination therapies, the gold standard of malaria treatment, which is emerging on the Thai-Cambodian border sooner than anyone expected (p. 844). Another piece examines how the research agenda has changed since Bill and Melinda Gates uttered the “E” word, eradication, in October 2007 (p. 843). High on this new agenda are transmission-blocking vaccines that can neutralize a mosquito’s ability to pass on the disease (p. 847). Also, since the Gateses’ call, a number of countries that have malaria relatively under control have declared their intent to eliminate the disease within their borders, a challenge whose difficulties and costs should not be underestimated (p. 849). The island of Hispaniola, which Haiti and the Dominican Republic uneasily share, provides a striking example (p. 850).

In a related Editorial, Bloom discusses the need for the United States to continue its investment in the Global Fund to Fight AIDS, Tuberculosis, and Malaria. In the end, it will be careful, well-funded R&D, combined with political will and strong health care systems, that will help to lessen the impact of these terrible diseases of poverty.

— STELLA HURTLEY, CAROLINE ASH, LESLIE ROBERTS

Tuberculosis & Malaria

CONTENTS

News

- 842 Redrawing Africa’s Malaria Map
As Challenges Change, So Does Science
- 844 Malaria’s Drug Miracle in Danger
If Artemisinin Drugs Fail, What’s Plan B?
- 847 The ‘Do Unto Others’
Malaria Vaccine
- 849 Shrinking the Malaria Map
From the Outside In
Elimination Meets Reality in Hispaniola

Reviews

- 852 Tuberculosis: What We Don’t Know
Can, and Does, Hurt Us
D. G. Russell et al.
- 856 The Population Dynamics and
Control of Tuberculosis
C. Dye and B. G. Williams
- 862 That Was Then But This Is Now:
Malaria Research in the Time of an
Eradication Agenda
S. H. I. Kappe et al.
- 866 The Selection Landscape
of Malaria Parasites
M. J. Mackinnon and K. Marsh

See also Editorial p. 791 and Podcast

Science

NEWS

Redrawing Africa's Malaria Map

Smaller countries on the hardest-hit continent have shown how aggressive malaria control can slash cases and deaths. Can the big ones follow?

Africa is the key battlefield in the fight against malaria—and in recent years, there have been major successes. Thanks to the widespread introduction of insecticide-treated bed nets, indoor spraying, and a new generation of drugs called artemisinin-based combination therapies (ACTs), a dozen countries have dramatically reduced their disease and death toll faster than anyone expected. “The number of cases and deaths

has fallen shockingly rapidly,” says Robert Newman, head of the World Health Organization’s Global Malaria Programme.

But most of the successful countries were small and peaceful. More than half of all malaria cases occur in Nigeria, the Democratic Republic of the Congo, Sudan, Tanzania, and Uganda—and in some of those, the challenges are far greater. Other questions loom as well. Can the countries that have

seen malaria dwindle keep up their vigilance? And will donors remain interested?

Looking at the map of Africa, however, malaria fighters say they can’t help but feel a little relieved. “There’s still a lot to do,” says Richard Steketee of the malaria control program at PATH in Ferney-Voltaire, France. “But we’re in a very different place than we were 10 years ago.”

—MARTIN ENSERINK



NIGERIA

Population: 155 million
2009 Funding: \$2.62*

Accounting for about a quarter of Africa’s malaria burden, this massive country is difficult for donors to deal with and money has been scarce. There’s been little progress in malaria control in the past few years. But things might change in 2010, when dozens of millions of bed nets will be deployed.

BED NET
COVERAGE
LESS THAN
10%

ETHIOPIA

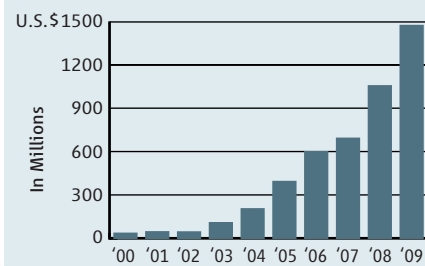
Population: 81 million
2009 Funding: \$7.94*

Ethiopia is the only big country to have introduced bed nets, ACTs, and indoor spraying on a large scale. Malaria here comes in epidemic waves, the last one in 2003–04, so the question is: How big—or small—will the next one be?

MOST
HOUSEHOLDS
NOW HAVE
NETS

A Funding Explosion

Donor contributions for malaria control to endemic countries worldwide.



SÃO TOMÉ AND PRÍNCIPE

Population: 0.16 million
2009 Funding: n/a

Everything is a bit easier if you’re a two-island state the size of New York City. Nevertheless, São Tomé and Príncipe, where malaria deaths have come down by more than 95% since 2005, is an example of what’s possible with existing tools.

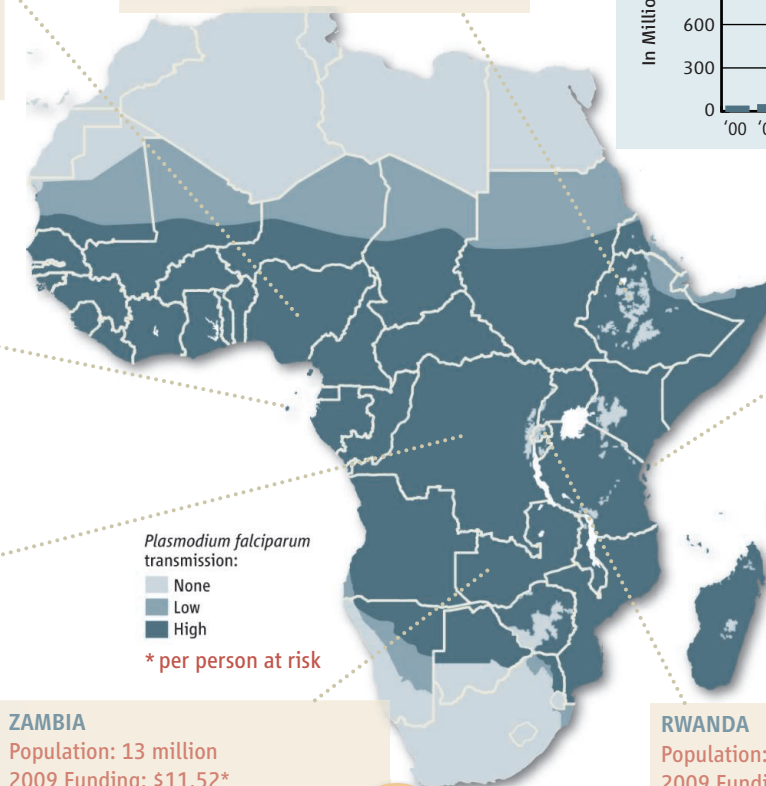
MALARIA
ALMOST
GONE

DEMOCRATIC REPUBLIC OF THE CONGO

Population: 69 million
2009 Funding: \$1.61*

One in nine African malaria cases occur in the D.R.C. War-torn, huge, and with a dismal infrastructure, experts say it’s one of the most difficult places anywhere to fight malaria. Recommended drugs are in short supply; nonetheless, 12 million bed nets were deployed between 2006 and 2008, and more are on the way.

MALARIA
STILL #1
KILLER



Plasmodium falciparum
transmission:

None
Low
High

* per person at risk

ZAMBIA

Population: 13 million
2009 Funding: \$11.52*

Zambia is a midsize “donor darling” that’s often held up as an example of what’s possible, and malaria deaths have come down by more than two-thirds since aggressive control measures were scaled up countrywide in 2003.

MALARIA
DEATHS
DOWN
65%

ZANZIBAR

Population: 1.1 million
2009 Funding: n/a

The small archipelago of Zanzibar—part of Tanzania—began seeing spectacular success in 2006 after 3 years of ramping up control measures. Overall childhood mortality was more than halved. Should Zanzibar consider going from control to elimination? (See p. 849.)

CHILDHOOD
MORTALITY
DOWN
52%

RWANDA

Population: 10 million
2009 Funding: \$30.74*

Among the front-runners in terms of funding, Rwanda rapidly introduced bed nets and ACTs nationwide in 2006. Malaria cases fell sharply in 2007; a recent study says in-patient deaths among children under 5 came down by 67%.

MALARIA
DOWN
MORE THAN
50%

As Challenges Change, So Does Science

Malaria scientists are brimming with optimism. Gone is the gloomy atmosphere that dominated the '90s, when nothing seemed to get done and funding was always in short supply. Thanks to a massive influx of new money, dozens of countries have begun scaling up simple ways to fight malaria: insecticide-treated bed nets, indoor spraying of insecticides, and a class of powerful drugs called artemisinin-based combination therapies (ACTs). Some have seen spectacular results. On top of that, megaphilanthropists Bill and Melinda Gates gave the field a breathtakingly ambitious goal in 2007: Don't just control malaria, eradicate it—get rid of it once and for all.

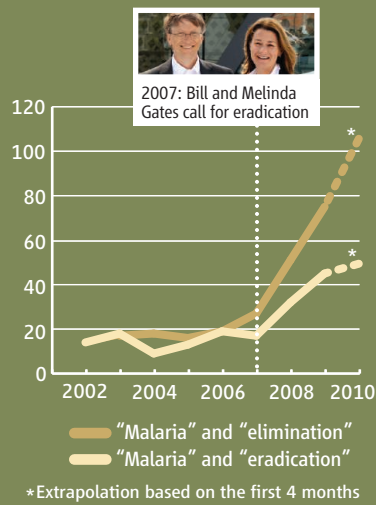
Nobody believes that is possible in the next 10, 20, or even 30 years. But the call for eradication, combined with the plummeting malaria burden, are

already reshaping the scientific agenda. Rather than preventing disease and death, there's a new focus on breaking the chain of transmission between host and parasite. Countries that have brought down cases by more than 90% are asking, "Now what?" Can they eliminate malaria—that is, get local transmission down to zero? The number of papers mentioning eradication and elimination in the context of malaria has exploded.

Over the past 18 months, a group of 15 scientists, supported by the Gates Foundation and led by Pedro Alonso of the Barcelona Centre for International Health Research in Spain, has consulted with hundreds of experts to hash out a new road map, called the Malaria Eradication Research Agenda (malERA). The fruits of their labor will be published in the months ahead. But based on interviews with malaria scientists and previous papers, here's a look at what's up and down on the research agenda.

—M.E.

Papers with "malaria" and "eradication" or "elimination" in the title or abstract



↑ Transmission-blocking drugs

Experience with ACTs has shown that drugs don't just save individual patients. They can also prevent others from getting sick, because treated patients become less infective to mosquitoes that bite them. But ACTs aren't ideal for this purpose, because they don't kill all the gametocytes, the cells in the parasite's sexual stage that get picked up by the mosquito. Another drug, primaquine, does a much better job, but it also causes the breakdown of red blood cells in patients with certain variants of a gene called *G6PD*. So safer gametocyte-killing drugs are needed, as well as research into the possibility of screening patients for their individual risk of primaquine side effects.



But the surprising successes achieved by existing cheap tools make these vaccines less interesting. "Maybe we won't need a

vaccine that does what bed nets can do," says Carlos Campbell, who directs the

malaria control program at PATH, a Seattle, Washington-based nonprofit.

↑ Transmission-blocking vaccines

The hunt is on for new generations of vaccines that don't protect the recipient from being infected but prevent mosquitoes from passing on the parasite to other people. Several of them are in development (see main text, p. 847).

↓ Partially effective vaccines

Developing a fully protective malaria vaccine has proven elusive. But because of the disease's terrible burden, researchers have long believed that even a partially effective vaccine can do a great deal of good, and one such vaccine—GlaxoSmithKline's RTS,S, which offers about 50% protection the first 8 months—is now in a phase III trial in seven African countries.

↓ Intermittent preventive treatment

In so-called intermittent preventive treatment (IPT), malaria drugs are given to people at high risk, such as children and pregnant women, in malaria-endemic areas to prevent them from becoming sick. That strategy becomes less attractive when malaria cases have been sharply reduced, because it means treating more and more

healthy people for one malaria case prevented. Indeed, some countries where malaria rates have plummeted are beginning to question the use of IPT in pregnant women. Instead, research is needed to find other ways to protect high-risk groups, scientists say.



Today, however, scientists say a few rounds of carefully planned mass drug administration, combined with other measures, might deal parasites the coup de grâce in countries on the brink of elimination. But it will take a new drug combination that is safe enough to give to hundreds of thousands of people and that, ideally, is so effective that a single dose will do. Another option would be to screen entire populations and treat only carriers of malaria. Modeling studies are needed to predict the effects of both strategies.

↑ Mass drug administration

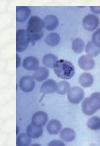
Treating entire populations with malaria drugs has been tried since the 1930s; chloroquine or amodiaquine was even added to salt in a dozen countries. The lat-



ter strategy is now widely seen as "stupid," says Brian Greenwood of the London School of Hygiene and Tropical Medicine; such massive use easily triggers resistance, and the spectacular successes in controlling malaria it brought were usually short-lived.

↑ *Plasmodium vivax*

Plasmodium falciparum is a global mass murderer. Its lesser-known cousin, *P. vivax*, found in Asia, South America, and Africa, is far less deadly. From the eradication viewpoint, however, *P. vivax* is a bigger problem because it can lurk in the human liver for years and come back unexpectedly. *P.*



vivax needs more study, researchers say—and new drugs are needed that kill the sleeper cells, called hypnozoites, in the liver. The Medicines for Malaria Venture has one called tafenoquine in a phase II trial.



NEWS

Malaria's Drug Miracle in Danger

Like many others before it, the latest generation of malaria drugs is losing its punch. This time, can global disaster be averted?

BANGKOK—Malaria rates are plummeting in many places, and scientists are optimistically talking about ridding entire countries of the disease—or even, in the long run, eradicating it worldwide. But in Southeast Asia, a new threat is looming—one that so far has received little attention but that could wipe out the recent advances and set back the global fight by decades.

Along the border between Thailand and Cambodia, *Plasmodium falciparum*, the most dangerous of malaria parasites, is showing unmistakable signs of becoming resistant to artemisinin derivatives, the group of powerful drugs that—as part of so-called artemisinin-based combination therapies (ACTs)—have become the cornerstone of malaria treatment worldwide.

For the moment, the drugs still work in the area; they just take more time to do the job. But that is alarming enough, scientists say. If resistance gets worse and starts spreading out from here, the results could be catastrophic. There are few other drugs in the pipeline, and those that are closest to the market may not work against the resistant strain (see sidebar, p. 846). One infected person flying to, say,

India or Africa could unleash the resistant parasites there.

Plans are under way to minimize the risk of that happening. The idea is to control malaria as aggressively as possible in the border area, making sure patients don't pass resistant parasites on to mosquitoes, and keeping parasites from being exposed to nonlethal drug levels that could fuel more resistance. Already, large numbers of insecticide-treated bed nets have been distributed in the area, and Cambodia has enacted a ban on the sale of so-called monotherapies, pills that contain artemisinin but lack the double whammy provided by the second drug in an ACT.

The World Health Organization (WHO), which coordinates the containment effort, bills it as unprecedented; when previous malaria drugs started becoming ineffective, the world sat by and watched it happen, says Robert Newman, head of WHO's Global Malaria Programme. But some malaria experts aren't satisfied. "There have been endless meetings but not enough action," says Nicholas White, who's widely considered the dean of malaria research in Asia. Because millions of lives are at stake, White, employed by the British

On the frontline. Chhay Meth, 9, suffers from malaria at her family's home in O'treng in Pailin, a province in western Cambodia where drugs are becoming less effective.

Wellcome Trust and the University of Oxford but stationed at Mahidol University in Bangkok since 1981, says the best strategy would be to eliminate malaria from the border area altogether. That idea, discussed widely a few years ago, has so far gone nowhere.

Malaria scientists who talked to *Science* were careful not to come down too hard on WHO, because the agency is chronically understaffed and because they want to see Newman—a respected malaria veteran who's been on the job less than a year—succeed. But they say the agency needs to step up its efforts, and the world should pay more attention. "Blame is not relevant, but outrage is entirely appropriate," says malariologist Christopher Plowe of the University of Maryland (UMD) School of Medicine in Baltimore. Given the huge stakes, White argues that the fight against resistance should "be prosecuted like a war. Because it is a war."

Déjà vu

If, a decade ago, you had asked malaria researchers where artemisinin resistance might first crop up, they probably would have put their money on the Thai-Cambodian border, a region that has been the cradle of resistance multiple times. In 1957, chloroquine,

discovered by Bayer in the 1930s and deployed widely after World War II, began failing here. Resistant *P. falciparum* started spreading in all directions in Asia, as well as to east and eventually west Africa. A separate resistant strain that arose along the Panamanian-Colombian border fanned out across South America, and by 2000, the drug was virtually useless in most parts of the world.

Chloroquine's main successor, sulfadoxine-pyrimethamine (SP), also known as Fansidar, started losing its potency here in the 1960s; this time resistance spread much faster. SP's replacement, mefloquine—developed by the U.S. Army during the Vietnam War and notorious among Western travelers for its side effects—started failing in Southeast Asia in the early 1980s. It's still effective in most other parts of the world.

Why the region is such a hotbed isn't clear, and the reasons are probably diverse. Widespread overuse of drugs may have played a role, as may the trade in counterfeit drugs. Fake drugs often contain lower levels of a drug—enough to make the patient feel better but too little to kill the parasites. (The same can happen when patients take low doses of a legitimate drug.) Some scientists have also speculated that the parasite population in this region may be more genetically diverse than elsewhere, giving it an evolutionary leg up. Whatever the causes, history appears to be repeating itself, sooner than anyone anticipated.

Hard to track

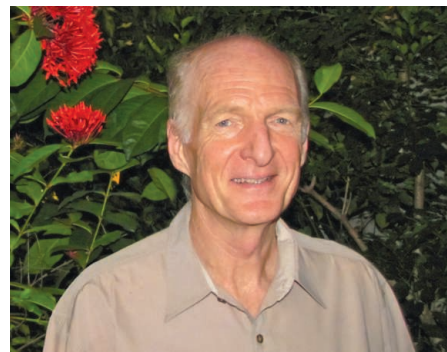
Chinese scientists isolated artemisinin in the 1970s from sweet wormwood (*Artemisia annua*), a plant used to treat fever for centuries. The drug, along with more potent derivatives such as artemether, artesunate, and dihydroartemisinin, not only kills parasites effectively and fast but also disappears from the bloodstream in less than a day after administration. The hope was that this would give the parasite less chance to adapt and develop resistance, and that teaming each of the drugs with another, longer-lasting one that mops up any remaining parasites would further reduce chances of resistance.

Perhaps that's why many didn't believe there was a seri-

ous problem when the Thai Ministry of Public Health first reported that treatment with the combination artesunate-mefloquine—the most-used ACT in Southeast Asia—occasionally failed in the province of Trat, which borders Cambodia, in 2003 and 2004. (Thailand is almost entirely malaria-free; the disease occurs only along its borders.) Most experts believed the problem was resistance against mefloquine, the partner drug, a well-known problem.

Fingering artemisinin wasn't easy, says Harald Noedl of the Medical University of Vienna, who carried out the pioneering resistance studies with researchers from the U.S. Armed Forces Research Institute of Medical Sciences in Bangkok. Scientists haven't yet found genetic markers by which they can easily identify resistant *P. falciparum* strains. The only surefire way to track artemisinin resistance is to treat patients with the drug—and without the partner drug—and check how fast they clear the parasites. WHO and ethical panels have concluded that this is ethically acceptable, says Noedl, provided subjects get a 7-day course of artesunate—instead of the standard 3 days for the combination—and they are followed carefully for up to 9 weeks. Such “treatment trials” are time-consuming and expensive, and they have been slow to get off the ground.

One study in western Thailand showed slightly increased parasite clearance times, but not nearly as dramatic as in Cambodia. It's impossible to say whether the change



General needed. Nicholas White says the fight against drug resistance “is a war.”

represents a very early stage of resistance; patients could also take longer to get rid of malaria because malaria control has been so effective that their immunity has waned, says François Nosten, who heads the Shoklo Malaria Research Unit in Mae Sot, a town on the Thai border with Myanmar. If resistance spills over into Myanmar, where the health infrastructure is poor and malaria is rife, most experts agree it would be impossible to stop it from spreading farther west to Bangladesh and India.

What's happening?

Action is needed, and the Thai and Cambodian governments, WHO, and a range of donors have all stepped up their efforts; the Bill and Melinda Gates Foundation alone has contributed \$22 million. What's been lacking, however, is a clear plan, a common strategy, and strong leadership, says White. “I'm in the epicenter here, and even I have trouble finding out what's happening,” he says. When *Science* asked WHO, which coordinates the efforts, about the current strategy, Newman said WHO has a plan but conceded that no publicly available version of it existed; he had WHO staff members write up a strategy and a progress report, however, which he sent to *Science* on 28 April.

Among the accomplishments listed in the report: Cambodia is improving its surveillance and has recruited more than 1300 volunteer malaria workers to help diagnose patients early and give them the right treatment. Bed nets are ubiquitous, and malaria

Cradle of Resistance



On the border. WHO's containment plan focuses on areas where resistance has been confirmed but includes a much larger region considered at risk as well.

Tuberculosis & Malaria

rates are dropping. Cambodia also banned monotherapies in 2009, and that seems to be working, at least in some places, says Arjen Dondorp, the head of malaria research in White's group; when Dondorp asked a Cambodian co-worker last year to see whether he could buy the drugs at local pharmacies in the province of Pailin, the man came back empty-handed.

But the challenges remain formidable. Western Cambodia has many migrants from

neighboring countries looking for work, and they're difficult to find and treat. Nobody even knows how many people are in the area, says Charles Delacollette, the coordinator of WHO's Mekong Malaria Programme. Bed nets don't work as well here as in Africa because mosquitoes start biting earlier in the evening, before people go to bed, and bite outside homes as well.

The tensions between Thailand and Cambodia—just a few weeks ago, their

armies briefly exchanged shots in a disputed border zone—don't help. One urgent question is whether a new, so-called synthetic artemisinin might work against resistant parasites. A study to find out was approved by ethical panels and ready to go when Cambodia's minister of health unexpectedly vetoed it; apparently part of the reason was that researchers from Thailand were involved, says Stephan Duparc, chief scientific officer of the Medicines for Malaria Venture, which supports the study. Still, experts from both sides of the border do meet frequently without political interference, says Wichai Satimai, director of Thailand's Bureau of Vector-Borne Disease.

The idea of wiping out malaria from the region altogether, meanwhile, has for now been shelved. On paper at least, the plan has potential. In terms of malaria, the area is almost an island, surrounded by non-malarious areas to the north, west, and east and by the Gulf of Thailand to the south. If the disease were eliminated, it might bounce back—but it would do so by being reimported from another region, such as eastern Cambodia, where there is no resistance.

White believes an approach called mass drug administration, in which everybody in the area receives drugs in a short time (see sidebar, p. 843), could do the job. "It would be a massive operation. You'd need global positioning systems, the military, and who knows what else to find all the people," he says, "but I think it could be done." But WHO is generally not in favor of mass drug administration, which targets people regardless of whether they have malaria and carries the risk of aggravating resistance if it's not carried out properly. WHO describes elimination as the "ultimate goal."

Newman resents the perception that the threat is not being taken seriously enough. WHO has lobbied hard against monotherapies, he points out, and WHO Director-General Margaret Chan has raised the issue of artemisinin resistance multiple times. But UMD's Plowe says that isn't nearly enough. "There should be a czar for malaria elimination in Southeast Asia, with a strong international team," he says. "Right now, I don't know anyone whose full-time job it is to contain resistance."

White agrees; the stakes are high enough, he says. "I would hate for resistance to show up in Mali next year and to know that we didn't stop it. I think we would all feel that we had failed terribly."

—MARTIN ENSERINK

If Artemisinin Drugs Fail, What's Plan B?

If the current generation of artemisinin-based combination therapies (ACTs) should fail, does the world have a solid backup plan? The short answer: No. New funding has lured academics and pharmaceutical companies back to the malaria field after a decades-long drought. But drug development is a slow process, and the only compounds far enough along in the pipeline to quickly replace current ACTs are variations on the artemisinin theme. Whether they will kill resistant parasites is a big question.

The three artemisinin derivatives currently used in ACTs are so closely related chemically that parasites resistant to one will probably be resistant to all, says Harald Noedl of the Medical University of Vienna. In other words, you can't swap one for the other. One possible alternative is a new artemisinin derivative, developed by the Hong Kong University of Science and Technology, whose structure is different from that of the others.

The drug, called artemesone, has been through a phase II trial for nonresistant malaria but didn't offer major benefits over existing derivatives, says Timothy Wells, chief scientific officer at the Medicines for Malaria Venture (MMV), a nonprofit that links scientists and companies to get new drugs to the market. It might have a new future, however, if the resistant parasites found along the Thai-Cambodian border prove sensitive to it. MMV is supporting a proposed trial in eastern Thailand to find out.

Then there are the synthetic artemisinins. These drugs, developed by researchers at the University of Nebraska, Monash University in Australia, and the Swiss Tropical Institute, were originally designed to circumvent the shortages and high prices of sweet wormwood, the plant from which artemisinin is isolated. Made from scratch in the lab, the molecules have the key chemical feature of artemisinin, a so-called endoperoxide bridge, but other than that, they are completely different.

MMV stopped sponsoring development of the first of these compounds, arterolane, after a disappointing phase II trial for nonresistant malaria, but it has renewed its collaboration with Indian pharma giant Ranbaxy to test it against resistant malaria, paired with a drug called piperaquine in an ACT. The three academic groups, meanwhile, have developed a second, more potent candidate, provisionally called OZ439, that will also be tried as soon as it has proved to be effective against nonresistant malaria in a phase II study.

Based on their different chemical structures, Wells feels chances are "pretty good" that the synthetics can replace standard artemisinins if they fail. Others are less sanguine because the mechanism of resistance is still unknown. There's a chance of success if, for instance, parasites become resistant through a pump that removes the drug from their cells, says Arjen Dondorp of Mahidol University in Bangkok; such pumps are unlikely to work for two molecules with distinct shapes. But if resistance occurs because the drug's target has changed a bit—making the fit less snug—then the new synthetics may not do any better.

Beyond artemesone, arterolane, and OZ439, there's not much else in sight. There are plenty of new leads, but none of them are even in phase I safety studies, Wells says, and they would take at least 7 or 8 years to get to the market.

—M.E.



No successors. ACTs—easy to take and fast-acting—can't be easily replaced.

NEWS

The 'Do Unto Others' Malaria Vaccine

Once neglected, research on transmission-blocking vaccines for malaria is gaining new prominence

It used to sound like a far-out idea: a malaria vaccine that would use humans to generate antibodies and deliver them to mosquitoes, with the aim of preventing the insects from spreading the disease. Until recently, such transmission-blocking vaccines (TBVs) had received scant attention. But after Bill and Melinda Gates called for a global effort to eradicate malaria (*Science*, 7 December 2007, p. 1544), TBVs have moved to center stage in malaria vaccine research (see p. 843).

With money no longer the limiting factor, progress is accelerating: A half-dozen TBVs could have a shot at clinical trials sometime in the next 5 years, researchers say. Even so, the scientific and social hurdles remain daunting. Key among them is whether people can be persuaded to get a vaccination that doesn't prevent them from getting sick but instead protects family and neighbors from getting infected. That also raises the bar for an extremely safe vaccine.

Nirbhay Kumar, a molecular parasitologist at Tulane University in New Orleans, Louisiana, has been working to develop a TBV for 28 years, often with little support. But he understands why few funders shared his enthusiasm: With limited resources, it made sense to focus on fighting the disease in humans. Consequently, the few candidate vaccines developed to date attack the parasite during the human stages of its life cycle. The frontrunner, called RTS,S, targets the parasite in the liver stage—when there are hundreds of parasites (see graphic). Other candidates target the blood stage, when there can be trillions of parasites. So far, such vaccines protect people from symptoms of malaria but don't clear all parasites from the blood.

As a result, a vaccinated person could be symptom-free but still be infected and infect others. "You can't eliminate or eradicate [malaria] without addressing the transmission" of the disease, says Kumar. "If we want to completely finish the job, we will need that vaccine," says Christian Loucq, head of the PATH Malaria Vaccine

Initiative (MVI), a nonprofit funded primarily by the Gates Foundation.

Playing the numbers

Antibodies generated by TBVs would be transferred to a mosquito during a blood meal, and they would zero in on the parasite after it enters the insect. That gives them a built-in advantage, says parasitologist Robert Sinden of Imperial College London: Instead of mopping up thousands or even trillions of parasites at once, they have to take out only a handful. The disadvantage? The vaccinated person can still get sick. But a mosquito that bites him or her won't pass the disease to others.

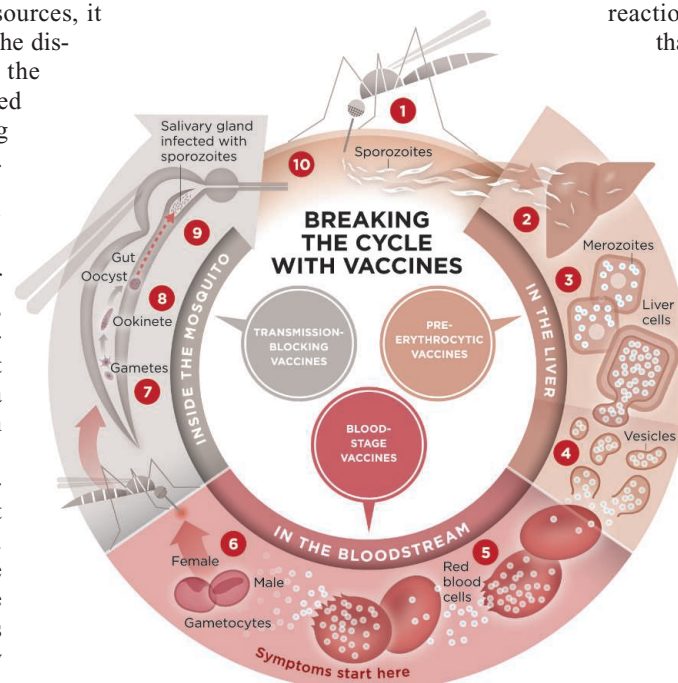
But even with the numbers advantage, developing this new class of vaccines isn't easy. TBV researchers not only have to aim at entirely new targets but they must also overcome many of the same problems that have plagued other malaria vaccines. One of those problems is that *Plasmodium* proteins—the targets of any type of malaria vaccine—are difficult to make in the lab. Usually vaccine producers use bacterial, yeast, or animal cells either to grow the target microbe

directly or to produce recombinant proteins, but the malaria parasite's genome can confound the proteinmaking machinery in these cells. To make matters worse, the antigens must have the right shape to trigger effective antibodies, and getting the recombinant *Plasmodium* proteins to fold correctly has proved difficult. Moreover, because TBVs won't provide direct benefit to the recipient, some say, any TBV with a chance in the real world has to be flawless: absolutely safe, 100% effective, and cheap.

A few teams began testing TBVs in humans a decade ago. The first antigen out of the gate was called Pfs25, named for a 25-kilodalton protein on the *Plasmodium falciparum* ookinete, the fertilized zygote that migrates through the mosquito midgut. Scientists have also tested the analogous protein from *P. vivax*, called Pvs25. (Researchers call both proteins P25 for short.) Like other *Plasmodium* proteins, P25 often failed to fold correctly when expressed in recombinant cells, but animal tests were promising enough to try it in humans.

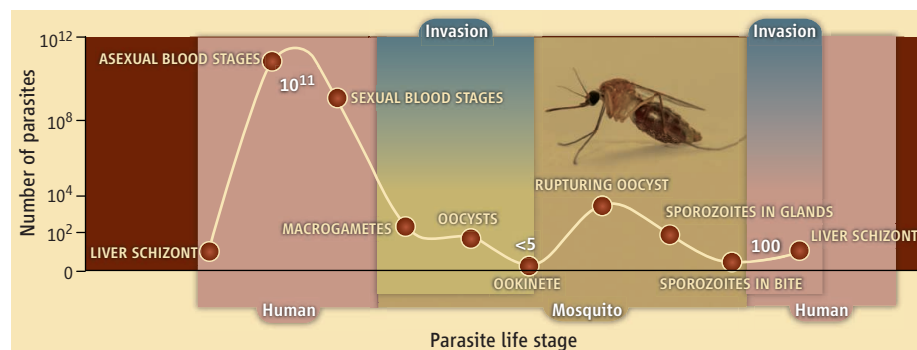
In initial clinical trials, vaccines based on P25 prompted measurable but weak reactions in healthy volunteers. Based on that experience, malaria experts Yimin Wu and Louis Miller of the National Institute of Allergy and Infectious Diseases in Rockville, Maryland, and their colleagues began a trial in 2005 that combined the antigens with an adjuvant that boosts the body's immune response to a vaccination. The researchers planned to test the vaccine in 72 healthy volunteers.

Initial results showed that some recipients produced antibodies that could block 90% of oocyst development in mosquitoes. But severe reactions in two participants stopped the trial. The combination of adjuvant and antigen was apparently too much for some immune systems, says Anna Durbin, a vaccine expert at the Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland, who coordinated the trial. Sev-



Multiple fronts. Different types of potential malaria vaccines target the parasite at different stages in its life cycle.

Tuberculosis & Malaria



Vulnerable target. The malaria parasite's population ranges from about five inside the mosquito to trillions in the human bloodstream.

eral groups are continuing work on P25 and the related P28, however, hoping that better folding might produce a potent response without new adjuvants.

At the same time, ongoing research on a half-dozen other candidates has brought them within range of clinical trials. One is Pfs48/45, about twice as big as P25 (and therefore harder to produce), which is expressed during the sexual stages of the parasite, both in the human and in the mosquito. Researchers identified it as a potential target and cloned the gene in 1993, "and we spent the next 15 years on its folding," says parasitologist Robert Sauerwein of Radboud University Nijmegen Medical Centre in the Netherlands. Eventually, several groups developed tricks to produce the protein in its lifelike shape. Kumar and his colleagues have conducted promising trials in baboons: The blood of animals that received a single immunization blocked parasite development by 93%.

Another candidate, called P230, is even more of a "monster," says Christine Farrance of the Fraunhofer USA Center for Molecular Biotechnology in Newark, Delaware. It is also expressed during the sexual-stage parasite, but it is 230 kilodaltons, too big to make in cell culture. Farrance and her colleagues are working with several teams to produce parts of P230 in tobacco plants, which are an inexpensive way to make complex proteins. Another leading candidate, called HAP2, is sometimes called the coitus interruptus antigen. First discovered in plants, but also present in *Plasmodium*, it is crucial for the fusion of male and female gametes. In lab tests, Sinden says, it can block parasite development by 96%.

One of the newest targets is particularly promising. Instead of targeting a parasite protein, the candidate vaccine targets a protein in the mosquito that is the docking point

for the parasite as it invades the insect's gut. Presumably, the antibodies bind to the gut protein, blocking the parasite's way and preventing its further development. One advantage to targeting a mosquito protein is that the antibodies are effective against several strains of parasite, says Rhoel Dinglasan, a malariologist at Johns Hopkins Malaria Research Institute in Baltimore. Antibodies from rabbits vaccinated with the antigen, called AnAPN1, blocked development of *P. vivax* by 98% in mosquitoes in Thailand and completely blocked *P. falciparum* development in mosquitoes in Cameroon, Dinglasan says. "It's the first time a candidate vaccine could do two with one blow," he says. The ability to hit both parasites with a single vaccine is a major advantage, says Ashley Birkett of MVI, which is collaborating with Dinglasan in one of its first forays into TBV research.

Clinical conundrum

As the various antigens move closer to human testing, experts are pondering how to design clinical trials. Vaccine clinical trials are never easy, but proving that a TBV can be effective in real life will pose new challenges, because the protection they offer is to communities, not to the vaccinated individual. Scientists have only begun to think about these issues, says epidemiologist Thomas Smith of the Swiss Tropical and Public Health Institute in Basel. "Large-scale trials of TBVs did not seem to be an imminent possibility until very recently," he says. And because recipients don't benefit directly, regulatory agencies may require even more safety data before approving them.

The epidemiology of malaria varies widely, and so vaccination strategies will differ from place to place. But the most likely scenario for an initial real-life test, Smith says, would involve several dozen communities with relatively low transmission. In half the communities, researchers would vaccinate as many people as possible with the TBV. The remaining villages would serve as controls. "The readout would be whether transmission was interrupted in the vaccinated communities," Smith says.

To get a handle on the factors that could affect a trial, Loucq says, MVI will hold a workshop in June that will bring together regulatory agencies, experts in mosquito behavior, epidemiologists, and statisticians who can help assess variables such as how far mosquitoes travel and how often they take a blood meal.



Lab test. Mosquitoes are fed on blood from vaccinated volunteers and *Plasmodium* parasites to test a vaccine's efficacy.

If a TBV proves effective in the field, the next step will be to see whether it could be combined with a vaccine that would prevent disease. "A multicomponent vaccine is what everyone is thinking about," Dinglasan says. If a transmission-blocking component could be added to RTS,S, Birkett says, "you would get the best of both worlds: a vaccine that has direct benefit to the recipient but also provides the community protection." At the same time, he says, any malaria vaccine has to be "very affordable, and combining antigens is much easier said than done."

A TBV will become increasingly important as other malaria-control measures have an effect and transmission falls, Sauerwein says. "It's very difficult to find that one case" that could lead to a new outbreak in an otherwise disease-free community, he says. "You need the vaccine to go the last mile."

—GRETCHEN VOGEL

NEWS

Shrinking the Malaria Map From the Outside In

A debate is brewing over how best to fight malaria: Go for the quick wins, or the worst places first, or both?

Richard Feachem wants to “shrink the malaria map.” By that, he and his Global Health Group at the University of California, San Francisco (UCSF), mean wiping out malaria at its “natural margins”—those countries on the edge of malaria transmission where the disease has just a tentative foothold—and working inward. It’s going for the “low-hanging fruit,” he says. “It’s a no-brainer.”

That’s not how some of Feachem’s colleagues see it. No one questions the value of eliminating malaria, which means stopping disease transmission in a geographically defined area—where it’s feasible today, although exactly what is feasible where and when, and what it might cost, is hotly debated. But many malaria hands worry that shrinking the map will divert money and attention from a far more urgent task:

reducing the horrific toll of malaria in central Africa, where five countries account for 50% of all global deaths from the disease and elimination is not possible. In those countries, scaling up the use of existing tools—for instance, blanketing the population with insecticide-treated bed nets and ensuring fast, free access to malaria drugs—could have a huge payoff. Indeed, several countries, such as Zambia and Rwanda, have slashed cases



No conflict.

Richard Feachem proposes a simultaneous attack on malaria’s strongholds and weak spots.

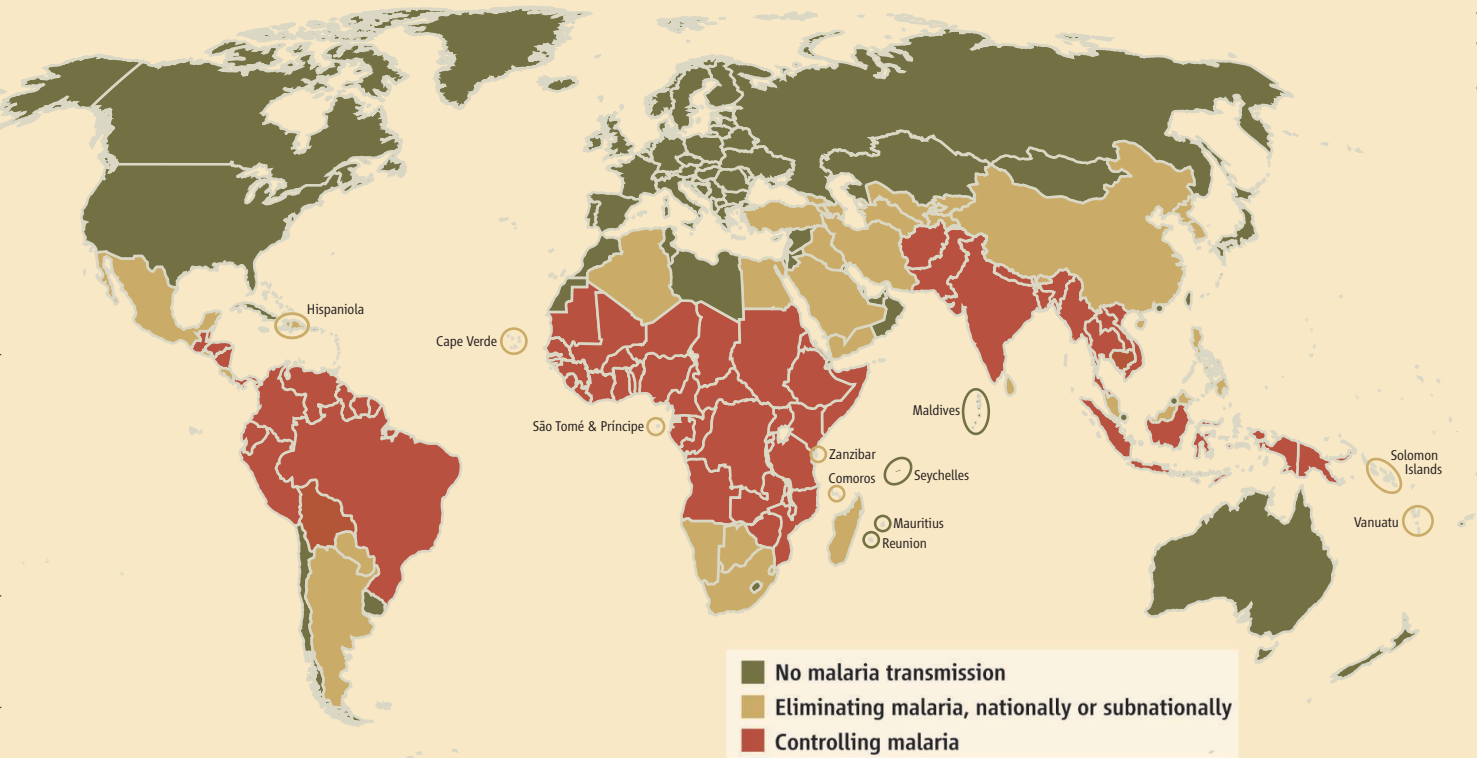
by more than 50% (see p. 842).

“There is not enough money for next year. So what investment would save the most lives and bring credibility to malaria control?” asks Carlos Campbell, who directs the malaria control program at PATH, a nonprofit in Seattle, Washington. “Do you work with Namibia and Botswana, which have very little malaria, and try to shrink the map, or do you use that money on bed nets for the DRC [Democratic Republic of the Congo]?”

Feachem, who has assembled a large brain trust known as the Malaria Elimination Group, agrees that the bulk of donor funds and efforts should be concentrated in the highest burden countries. But he thinks the world can and should pursue both strategies. There is “absolutely no conflict” between the two, he says. He sees elimination as a piece of a three-part strategy that must be pursued simultaneously for “getting to a malaria-free world”: Control malaria in the “heartland,” eliminate it at the edges, and invest heavily in radical new tools that someday might make it possible to wipe the disease off the face of the earth.

Many countries have already eliminated malaria—Canada and the United States, for example, and all of Western Europe. Several

CREDITS (TOP TO BOTTOM): COURTESY GLOBAL HEALTH GROUP; MAP ADAPTED FROM GLOBAL HEALTH GROUP



Tuberculosis & Malaria

others are in the process, including Algeria, China, and Mexico, all on the northern boundary of transmission, and Argentina and the South Pacific island of Vanuatu on the southern boundary. In these places, the environment is less conducive to malaria transmission. At least as important, the countries tend to be high- or middle-income, with sufficient funds, political will, and strong enough health systems to devote to chasing the last few malaria cases and ensuring the disease doesn't come back.

Feachem and the Malaria Elimination Group have published a 200-page prospectus, *Shrinking the Malaria Map*, that is essentially a how-to manual for getting local transmission to zero. In it they optimistically identify 39 countries as "eliminators" on the margins (see map, p. 849). Some of them, such as Haiti and the Domin-

ican Republic, will be at best an enormous challenge (see sidebar, below).

Getting from low to zero transmission anywhere is a big leap, Feachem says. In addition to continuing high-level control with existing tools, it requires an exquisitely sensitive surveillance system to detect and treat every case and monitor for the parasite's return once it has been eliminated. Reintroduction is a huge threat—success totally depends on your neighbors, explains Feachem. That's why islands are considered the lowest of the low-hanging fruit.

Four countries at the southern tip of Africa, where malaria transmission is now low, illustrate the challenges. South Africa, Swaziland, Botswana, and Namibia have declared their intention to eliminate malaria by 2015. The ambitious goal has been met with "skepticism," says Oliver Sabot of

the Clinton Health Access Initiative, but CHAI, in partnership with the UCSF Global Health Group, is helping those four countries achieve it. South Africa and Swaziland are making steady progress, Sabot says; Namibia and Botswana less so. Namibia boasts a strong health care system and good malaria control, he says, but it shares a border with Angola, which "has virtually no health care system and no control, and there is nothing to stop malaria from flowing across the border."

Robert Snow, who heads the Malaria Public Health and Epidemiology Group in Nairobi, is one of the skeptics. "Elimination costs a huge amount of money. Some of those are very poor countries. Is it an appropriate goal given the other health problems they face?" he asks.

Sabot contends that shrinking the map

Elimination Meets Reality in Hispaniola

There is absolutely no reason for malaria to persist on the island of Hispaniola, says Donald Hopkins, longtime disease fighter and vice president for health programs at the Carter Center in Atlanta. All the other islands in the Caribbean have rid themselves of this mosquito-borne disease. And the Dominican Republic (D.R.), which shares the island with Haiti, has driven cases to remarkably low levels. By many counts, the country is well on the path to elimination, its stated aim. But it can't get there unless Haiti, where malaria is rife, achieves the same goal. Hispaniola's fledgling program to eliminate malaria, now sidetracked by the earthquake, illustrates the difficulties that arise when ambitious goals collide with reality.

In 2006, Haiti reported 33,000 confirmed cases of malaria; experts at the U.S. Centers for Disease Control and Prevention (CDC) peg the number at closer to 200,000 and warn that it may surge as the million or so people displaced by the 12 January earthquake remain in crowded camps with no protection from mosquitoes and the rainy season upon them. The D.R., by contrast, reports about 3000 cases a year, a number believed to be fairly accurate, and those are found mostly along the border with Haiti and among migrants from that impoverished country. Still, the D.R.'s success in slashing cases—and the successes of the neighboring islands in wiping out the disease—show that malaria elimination on Hispaniola is technically feasible.

The biology is certainly obliging. Hispaniola is one of the few places in the world where the malaria parasite is still susceptible to chloroquine, a cheap and safe drug. Just one parasite, *Plasmodium falciparum*, transmits malaria on the island, as opposed to several in Africa, and although it causes the most lethal form of malaria, it is relatively susceptible to treatment.

There's just one vector, too—*Anopheles albimanus*, or "white hands," because of its distinctive white legs—and it is not very good at transmission, explains Angel Solís, an entomologist working with the Dominican health ministry. Unlike *A. gambiae*, the main vector in Africa, *albimanus* does not preferentially bite humans; a cow makes a fine meal. Nor does it prefer to bite indoors, where humans congregate—all factors that contribute to relatively low transmission there compared with Africa.

Finally, as an island, Hispaniola is protected by vast expanses of ocean,



Reaching out. Community health worker Jonel Mompremier pricks a boy's finger to test his blood for malaria parasites in Ouanaminthe, Haiti.

with no other malaria-endemic countries nearby. So if the parasite can be eliminated, chances are low that it will get back in.

Richard Feachem and colleagues at the University of California, San Francisco, have flagged Hispaniola as an "eliminator" on their malaria map (see main text, p. 849). And in 2006, the International Task Force for Disease Eradication (ITFDE), an expert group Hopkins chairs at the Carter Center, declared malaria elimination from Hispaniola "technically feasible, medically desirable, and ... economically beneficial."

But there is more to elimination than favorable biology, says Robert Newman, head of the Global Malaria Programme at the World Health Organization. A country or region must have a strong health system, lots of money, and a government or donors willing to invest in surveillance long after the disease has disappeared. "All the stars must be aligned," he says. Even before the earthquake, Haiti's malaria program, and even its public health system, essentially existed in name only. "There is a lack of nearly

gives countries that have already controlled malaria a much-needed goal: “The whole idea of shrinking the map is to give them a future to look to.” But even he worries that some countries, such as Ghana and Nigeria, are jumping on the elimination bandwagon prematurely, “without an awareness of what it actually means” either technically or financially. “From what we see so far, the expectation that a high-burden country can dramatically reduce funding for malaria once elimination is achieved is often false.”

Everyone agrees that “before going down the chute” of elimination, each country should conduct a rigorous feasibility assessment, says Robert Newman, head of the Global Malaria Programme at the World Health Organization (WHO). Unfortunately, he says, the data needed are sorely lacking. He and other experts laud Feachem’s group

and CHAI for leading the way in collecting them, and they point to a recent analysis CHAI and others did with the government of Zanzibar as the gold standard.

Zanzibar is one of the poster children for aggressive malaria control. With donor help, the islands massively scaled up interventions and reduced transmission dramatically. Now the government is grappling with what to do next, says Sabot: concentrate on sustaining those gains or try to stop transmission on the islands altogether?

The CHAI analysis concluded that malaria elimination in Zanzibar is technically feasible but would be “very challenging” both operationally and financially. “It would likely be substantially more costly than maintaining control” for the next 20 years, Sabot says. What tipped the scales is the risk of reintroduction. Although it’s

surrounded by water, Zanzibar is very close to mainland Tanzania, where malaria transmission is robust, and “there’s a ferry every hour,” says Newman. And once the parasite got in, it would quickly resurge, given the efficiency of the vector there.

“Many of us were surprised the feasibility assessment said that it is not practical in the near term,” says Newman. “And if an island setting shows elimination is challenging now, it will be challenging in other areas.” Newman says WHO plans to work with the Malaria Elimination Group to develop evidence-based guidelines to help countries decide whether to move from control to elimination. “Data and assessment should drive public health policy. You don’t want to set policy on enthusiasm, but you don’t want to rain on it either.”

—LESLIE ROBERTS

everything in Haiti,” says Nguyen-Dinh Phuc, a malariologist recently retired from CDC who has worked extensively on the island. “For malaria, they don’t have enough people trained in microscopy; they don’t have Ph.D. entomologists, or epidemiologists with expertise in malaria surveillance.”

Even so, after ITFDE’s proclamation, the Carter Center decided to see if it could prod the two countries to collaborate to eliminate malaria and another mosquito-borne disease, lymphatic filariasis (*Science*, 5 February, p. 634). The odds were against them, as elimination requires cross-border cooperation. “There was a lot of wariness between the two countries,” says Hopkins. “The impetus to collaborate was not strong.”

But the benefits far exceeded the risk of failure, Hopkins says. With a small amount of money and a considerable amount of technical expertise, the Carter Center helped launch a pilot project in 2008 in the border towns of Dajabón, D.R., and Ouanaminthe, Haiti, which straddle the infamous Massacre River. The location, the site of one of the bloodiest encounters between the two countries, was chosen deliberately: Not only is it a hot spot of transmission, but experts reasoned if the two could collaborate there, they could anywhere.

As part of the pilot project, Dominican malaria experts have been giving their Haitian counterparts a crash course in the malaria-fighting techniques the D.R. has long employed. They include active surveillance to track down every single malaria case; the replacement of presumptive treatment, in which any person with fever is treated for malaria, with confirmed laboratory diagnosis; and fast and free treatment with chloroquine to address the illness and primaquine to help reduce disease transmission. The campaign also includes

vector-control measures, such as bed nets and insecticide spraying, and monitoring for any signs of drug resistance.

Within a year and a half, the project, led by David Joa Espinal on the Dominican side and Joanel Mondestin on the Haitian, had trained 10 community health workers, distributed 1200 bed nets, hired and trained three microscopists, and bought four microscopes, 11 motorbikes, two computers, and 37 cell phones.

It was such a success, albeit a modest one, that in October 2009, with former President Jimmy Carter at their side, the health minister of each nation unveiled a binational plan to scale up these efforts across the entire island. The price tag: \$194 million over 10 years. Both presidents endorsed the plan, and Carter vowed to lobby the donor community for support.

Then the earthquake hit, and more urgent needs claimed priority. But as Haiti recovers, Hopkins and colleagues hope the idea of eliminating malaria and lymphatic filariasis can be revived.

The first order of business is to build a functioning public health system, with malaria control as a key component, says Larry Slutsker, who heads the malaria branch at CDC. Since the earthquake and several malaria outbreaks, CDC has had a team of malariologists on the ground to try to establish the rudiments of a control program in Port-au-Prince.

“Let’s keep the elimination goal—it is great to have a plan, and benchmarks—but let’s be realistic about what needs to be done first,” says Slutsker.

—L.R.

Leslie Roberts visited Dajabón and Ouanaminthe on a reporting trip in October 2009.



Easy access. Malaria hitches a ride into the D.R. with the throngs of Haitians who cross the Massacre River on market day.

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

The following resources related to this article are available online at www.sciencemag.org (this information is current as of May 13, 2010):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:
<http://www.sciencemag.org/cgi/content/full/328/5980/852>

This article **cites 61 articles**, 19 of which can be accessed for free:
<http://www.sciencemag.org/cgi/content/full/328/5980/852#otherarticles>

This article appears in the following **subject collections**:
Medicine, Diseases
<http://www.sciencemag.org/cgi/collection/medicine>
Microbiology
<http://www.sciencemag.org/cgi/collection/microbio>

Tuberculosis: What We Don't Know Can, and Does, Hurt Us

David G. Russell,^{1*} Clifton E. Barry 3rd,² JoAnne L. Flynn³

Mycobacterium tuberculosis has a penetrance of its host population that would be the envy of most human pathogens. About one-third of the human population would have a positive skin test for the infection and is thus thought to harbor the bacterium. Globally, 22 “high-burden” countries account for more than 80% of the active tuberculosis cases in the world, which shows the inequitable distribution of the disease. There is no effective vaccine against infection, and current drug therapies are fraught with problems, predominantly because of the protracted nature of the treatment and the increasing occurrence of drug resistance. Here we focus on the biology of the host-pathogen interaction and discuss new and evolving strategies for intervention.

Tuberculosis (TB) is a disease that, in the Western world, is held in check by effective public health systems that compensate for the relative shortcomings of current intervention strategies. However, because of the lack of vaccines and the need for new, faster-acting drugs, it is unclear how the disease can ever be controlled in the countries where it is truly endemic. In addition, as highly drug-resistant strains continue to evolve, we face the risk of losing control even in the industrialized world. Despite recent increases in research activity, we remain hampered by large gaps in our knowledge of the biology of this pathogen (Table 1). An increased appreciation of the bottlenecks in the life cycle of *Mycobacterium tuberculosis* (*Mtb*) should facilitate development of new intervention strategies that are applicable to those countries most in need.

The Life Cycle of *M. tuberculosis*

Infection with *Mtb* follows a pattern of events that have been established through animal models, as well as observations from human TB (1, 2) (Fig. 1). The infectious bacilli are inhaled as droplet nuclei that have been exhaled into the atmosphere. These droplets are small enough to remain airborne for several hours. Estimations of the minimum infectious dose range from a single bacterium upward. Our understanding of the initial stages of infection in the lung is mainly through inference; it is generally believed that the bacteria are phagocytosed by alveolar

macrophages, which then invade the subtending epithelial layer. This induces a localized inflammatory response that leads to recruitment of mononuclear cells from neighboring blood vessels, providing fresh host cells for the expanding bacterial population. These cells are the building blocks of the granuloma, which is the defining pathologic feature of this disease. Initially the granuloma is an amorphous mass of macrophages, monocytes, and neutrophils; however, the macrophages differentiate into several specialized cell types, including multinucleated giant cells, foamy macrophages, and epithelioid macrophages. With the development of an acquired immune response and the arrival of lymphocytes, the granuloma acquires a more organized, stratified structure. The macrophage-rich center becomes surrounded by a mantle

of lymphocytes that may be enclosed within a fibrous cuff that marks the periphery of the structure.

The appearance of *Mtb*-specific lymphocytes about 2 to 3 weeks postinfection marks the end of the phase of rapid bacterial replication and the onset of a “containment” state that, in mice, is characterized by relatively stable bacterial numbers. At this time, the granuloma is extensively vascularized, and cells are actively recruited to the site of infection. In granulomas exhibiting the pathology associated with disease progression, the fibrous sheath becomes more marked, and the number of blood vessels that penetrate the structure diminishes. There is also an increase in the number of foamy macrophages, which may be responsible for the increase in caseous debris in the center of the granuloma (3). At these “late” stages, the caseous portion of the granuloma becomes hypoxic (4), a condition that can induce a state of non-replicative persistence in *Mtb* in culture. Histology of infected tissues from immunocompetent patients with active TB reveals granulomas in all states of development from containment to active disease, which implies that the fate of each granuloma is determined locally, not systemically. Active granulomas exhibit extensive pathology, and, ultimately, the granuloma ruptures and spills thousands of viable, infectious bacilli into the airways (5), which results in the development of a productive cough that facilitates aerosol spread of infectious bacilli.

Recent observations in human TB patients indicate that neutrophil influx in late-stage disease may also contribute to the tissue damage and the dissemination of infectious bacteria into the air-

Table 1. Problems with prevention of and treatment for *Mtb* and shortcomings of current intervention strategies.

Global issue	Underlying regional problems
Failure to provide consistent infrastructure support for health care	- Inadequate public health sector in many regions
Failure to develop effective diagnostic and disease status indicators (biomarkers)	- Variation in disease manifestations across individuals
	- Inadequate testing for drug susceptibility and resistance
	- Reliance of current diagnostics mostly on the host immune response
	- Poor correlation of bacterial load in sputum with disease status or outcome
	- Lack of any indicators of lung pathology or health
Failure of bacillus Calmette-Guérin (BCG) vaccination to protect adult pulmonary TB	- Variability in host response to vaccination
	- Variability between BCG vaccine strains
	- Variability in bacterial strains causing infection
	- Variability in host genetics and/or the environment
Failure of chemotherapy to cure many patients with TB	- Length of therapy makes compliance difficult
	- Individual lesion types respond at different rates
	- Drug-tolerant bacteria develop subpopulations
	- New drug-resistant strains emerge frequently
Failure of target-based drug discovery programs	- Antibiotic killing involves complex biological processes
	- Genetic essentiality is not related to target vulnerability
	- Whole-cell screenings are not performed under conditions relevant to in vivo infection

¹Department of Microbiology and Immunology, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853, USA. ²Tuberculosis Research Section, Laboratory of Clinical Infectious Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, 33 North Drive, Building 33, Room 2W20D, Bethesda, MD 20892, USA. ³Department of Microbiology and Molecular Genetics, University of Pittsburgh School of Medicine, W1144 Biomedical Science Tower, Pittsburgh, PA 15261, USA.

*To whom correspondence should be addressed. E-mail: dgr8@cornell.edu

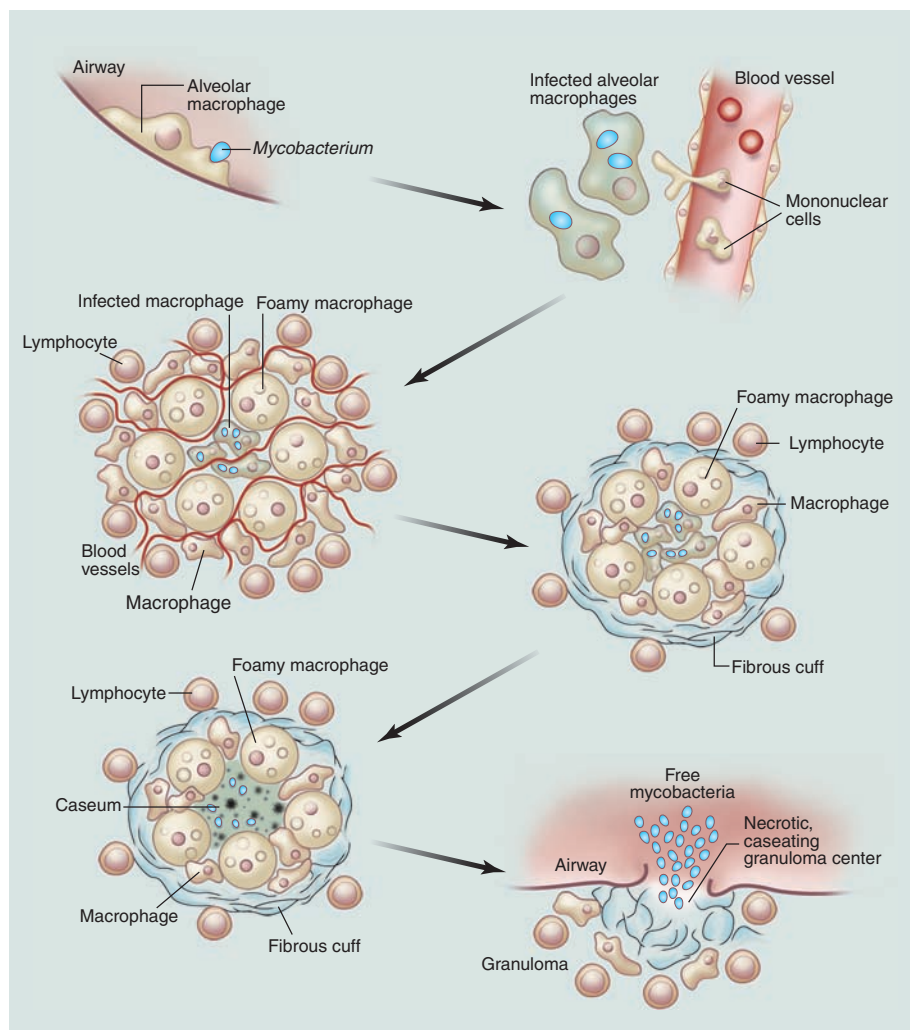


Fig. 1. The life cycle of *M. tuberculosis*. The infection is initiated when *Mtb* bacilli, present in exhaled droplets or nuclei, are inhaled and phagocytosed by resident alveolar macrophages. The resulting proinflammatory response triggers the infected cells to invade the subtending epithelium. This response also leads to the recruitment of monocytes from the circulation, as well as extensive neovascularization of the infection site. The macrophages in the granulomas differentiate to form epithelioid cells, multinucleate giant cells, and foam cells filled with lipid droplets. The granuloma can become further stratified by the formation of a fibrous cuff of extracellular matrix material that is laid down outside the macrophage layer. Lymphocytes appear to be restricted primarily to this peripheral area. Many of the granulomas persist in this balanced state, but progression toward disease is characterized by the loss of vascularization, increased necrosis, and the accumulation of caseum in the granuloma center. Ultimately, infectious bacilli are released into the airways when the granuloma cavitates and collapses into the lungs. [Adapted with permission from Macmillan Publishers Ltd. (3)]

ways (6). Modern imaging observations on human TB stress that the balance between containment and disease progression is complex and highly dynamic and appears to be a local phenomenon involving differential progression of individual granulomas within a single individual (7).

Immune Protection Through Vaccination

A preexisting, specific immune response against *Mtb*, acquired through vaccination or a chemotherapeutically resolved infection, has an impact on the course of TB (8) (Fig. 2). The immune

host is not protected against infection, but progression to containment occurs earlier, and at a lower bacterial load. In most animal models, bacterial containment is achieved with 1/10th as many bacteria as would be found in the lungs of a naïve animal. At the human population level, this would make a substantial difference, because progression to disease is linked directly to bacterial load. If disease is determined at the level of the individual granuloma, then the presence of fewer granulomas may reduce the chance of developing active disease.

Although the contribution of antimicrobial effectors has been established in mice (9, 10), the relative hierarchy of immune-mediated killing mechanisms in humans is unclear. However, from the increased susceptibility of HIV⁺ humans, we infer that CD4 T cells are important in the control of human TB. Similarly, the use of tumor necrosis factor (TNF)-neutralizing agents for treatment of inflammatory diseases substantially increases the risk of TB, which suggests that the level of this cytokine is critical to the balance between disease control and pathology (11–14). From these data and studies of genetic mutations that predispose humans to TB (15), we infer that, similarly to mice, macrophage activation in humans is central to the control of infection.

BCG: It's Not for Everyone

Bacillus Calmette-Guérin (BCG) is the only approved vaccine against TB. It was developed through the serial in vitro passage of *M. bovis* until it became nonpathogenic. It is used in countries with endemic TB because it protects children against severe forms of disease, such as TB meningitis or disseminated infection. However, although effective against development of TB in some countries such as the United Kingdom (16), its efficacy has been questioned in several studies, most notably in India, where very limited (or no) protection has been reported (17). There are three main hypotheses as to why BCG works in some populations but not in others. First, BCG has become too attenuated through culture, and modern preparations of the vaccine are too benign to generate adequate protective immunity (18). Second, exposure of infants to environmental mycobacteria in countries like India could lead to tolerance (19–21) or, third, clearance of the BCG in some populations may occur before development of a protective immune response. Clearly, as we move forward with new vaccine constructs, it is vital that we better understand the limitations of BCG-induced protection, so that a new vaccine can be effective in those countries where it is most needed.

New anti-TB vaccination strategies can be divided into three broad categories. First, improving BCG by adding or overexpressing strongly immunogenic *Mtb* antigens, which would enhance and broaden the immune responses induced by the recombinant bacterium (22–24). Second, attenuating strains of *Mtb* through the deletion of genes for specific metabolic pathways required for survival or full virulence (25–27). And third, the use of prime-boost strategies that direct and amplify an initial “protective” immune response through subsequent inoculation with viral vectors encoding *Mtb* antigens or protein subunits (28–33).

However, if the protective immune response to these strains is subject to the same environ-

mental factors that limit the effectiveness of BCG in humans, then we have not progressed. This may, however, be circumvented through the prime-boost strategy of supplementing vaccination with BCG, or recombinant strains, with nonreplicating adenovirus or vaccinia virus strains encoding *Mtb* antigens (28–33), which could supplement existing BCG vaccination programs, although this too has met with mixed results (30, 32).

Transition into the Field

So, given the limited efficacy of immune protection, where would vaccines be most useful in combating TB? Vaccination could play an important role in reducing transmission in areas that have a low incidence of HIV, particularly if there are multidrug resistant (MDR) strains of

changing landscape of circulating hypervirulent strains, such as the immunosuppressive Beijing lineage, which may have emerged in response to BCG vaccination, also interjects some caution into extrapolation from results generated in the laboratory (37, 38).

Chemotherapeutic Intervention

Current chemotherapeutic regimens require more than 6 months' treatment with multiple drugs. Such regimens are plagued with issues of patient noncompliance, inadequate health-care oversight, and the increasing proliferation of drug-resistant strains. The need for fast-acting, effective medications for use in resource-poor countries is immediate.

The balance of the physiological state(s) of the bacterial population inside the granuloma during the dynamic alterations in its structure is clearly central to persistence and progression to disease. Research involving bacterial strains defective in genes critical to various metabolic pathways has revealed different defects during the course of the infection cycle in mice. This implies, not surprisingly, that the bacterium regulates its metabolism differentially during the progression of disease (7). Such metabolic shifts could be driven by the availability of certain nutrients or carbon sources, the aerobic or hypoxic nature of the granuloma, or the level of stress induced by various infected host cells and the immune response. Because of the variability between granulomas within a single infected individual and their highly differentiated internal structure, granulomas offer an extremely diverse range of environments for *Mtb*.

In a contained lesion, the bacterial load appears relatively static, and it has been hypothesized that the bacteria are in a nonreplicative state of vegetative metabolism (39, 40). However, more recent work monitoring bacterial replication via an unstable “clock” plasmid indicates that the bacterial population, even in apparently stable lesions, continues to undergo replication (41). It is unlikely that there is a simple binary distribution of bacteria between replicating and nonreplicative status, and even within a single granuloma, there are likely to be multiple local microenvironments supporting unique bacterial populations. The drugs that are currently in common usage preferentially target replicating organisms; therefore, a nonreplicative subpopulation of the bacteria present in the persistent or “latent” infection will show innate resistance to drugs. The innate resistance of nonreplicating *Mtb* may contribute

substantially to the protracted treatment period required in current therapeutic regimens.

Bacterial Metabolism During Infection

Although the murine model for TB is imperfect, it has proven to be an extremely useful tool in probing some of the metabolic shifts required to sustain a bacterium transitioning from rich broth into an in vivo infection (42, 43). One good example of a metabolic shift required to support infection is the realignment of the bacterium's lipid metabolism. Isocitrate lyase activity is required to sustain *Mtb* infection in the chronic phase of infection in the murine model (44). The enzyme appears to fulfill a detoxification function, controlling propionate levels when the bacterium uses specific carbon sources, such as cholesterol (45–49), which has been invoked as a nutrient during the persistent phase of infection (46, 50).

Recent studies with the hypoxia indicator pimonidazole, coupled with direct measurements with an oxygen probe in lesions of primate and rabbit models, have confirmed previous speculation that the inner regions of a fully stratified granuloma have extremely low, but detectable, oxygen tension (4). This has implications for tissue pathology, but it may also affect the metabolic state of the bacterium, which up-regulates the DosR regulon to promote survival and re-emergence into the growth state (51).

Drug Discovery Programs: Introduction of Realism at the Level of Screening

Drug discovery programs for the identification of novel targets, as well as new antimicrobials, are few and far between. The more traditional methods of target-based discovery and remodeling of recognized scaffolds have pretty much exhausted known possibilities. The search for anti-infectives, as a field, has placed an inappropriate emphasis on the need to demonstrate genetic “essentiality” for any target, which is based on the misplaced assumption that genetic knockouts can be phenotypically copied by small-molecule inhibitors of metabolic pathways. We need to shift our mind-set from the discovery of essential gene products to appreciating the subtle holistic interactions between pathways and gene products that are required for survival in the many different microenvironments of the host.

Increasingly sophisticated cell-based screens of *Mtb* in a variety of conditions that simulate the multiple microenvironments of the host are needed to identify those metabolic breakpoints that could feed new target-based development programs. From both the basic and the translational science aspects, understanding the adaptation of the organism and its points of vulnerability under various environmental conditions represents a logical pathway toward agents potent within the granuloma that may shorten therapy. Just as important is a better un-

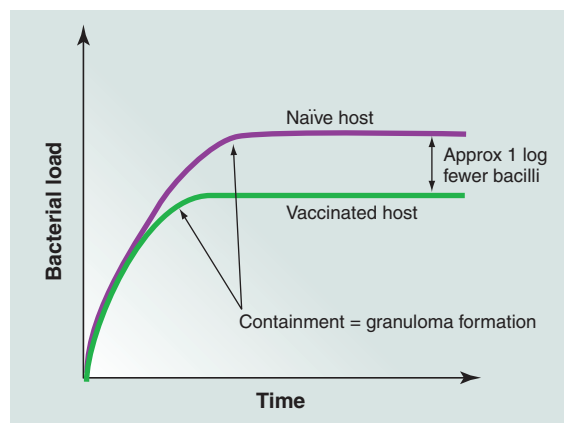


Fig. 2. The bacterial load in naïve and immune hosts. The course of *Mtb* infection in naïve and immune hosts follows a reproducible pattern in the mouse model. After infection, the bacteria replicate exponentially for a period of time until the host develops an acquired immune response, which limits bacterial replication. At this time, the infection makes a transition to a state of persistence or chronicity where the bacterial load is sustained at a relatively constant level. Transition into this persistent state occurs earlier in a vaccinated host or a host that has been infected and treated chemotherapeutically. In mice vaccinated with BCG, this transition occurs with a bacterial load that is about 1/10th that of a naïve host. The transition into persistence is marked by the development of a granuloma.

Mtb present. However, the impact of vaccination on populations with high incidence of HIV, such as sub-Saharan Africa, is open to question. Transmission of TB appears to be less efficient in HIV patients because a robust immune response is a major contributor to the pathology required for transmission (34–36). However, in areas where the coincidence of HIV and *Mtb* infections exceeds 80% of TB cases, this is unlikely to have a discernible impact. We have yet to see if these new modified vaccine strains will improve the protection afforded by BCG or will recapitulate the absence of efficacy of BCG vaccination among ethnic groups most at risk from the disease. The

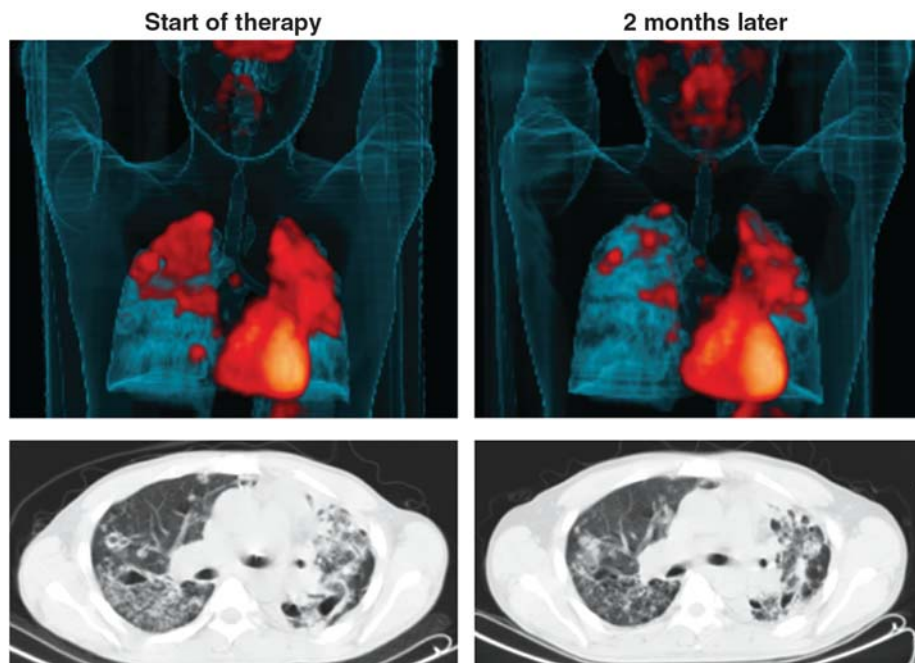


Fig. 3. Imaging disease progression in humans. 18-Fluorodeoxyglucose (FDG) PET and CAT scans from a patient with pulmonary TB: **(left)** at the time when therapy was initiated and **(right)** 2 months after treatment has begun. Below are two matched CAT slices from the area indicated by the plane on the left taken from the same scans. FDG is primarily a marker of increased tissue metabolism, such as inflammation, whereas CAT shows the structural abnormalities within this patient's lungs. Both sets of images illustrate the wide range of lesion types occupied by *Mtb* during infection that show differential kinetics of response to chemotherapy.

derstanding of the microenvironments experienced by the bacterium in the human host, so that these can be replicated *in vitro*. As one example, screening directly on *Mtb*-infected macrophages in differing states of activation would incorporate many of the host-mediated pressures responsible for inducing the different replicative and nonreplicative states of *Mtb* observed in the human host (52).

Missing Tools

As the field moves forward to evaluate new drugs and vaccines, we are clearly lacking some of the most basic tools for assessing efficacy. The mouse model fails to form granulomas that reproduce the tissue environments seen in humans. Accurate reproduction of these environments is likely to be important to both chemo- and immunotherapy. Rabbits and guinea pigs generate a tissue response that appears more comparable to that of humans, but it still requires validation in both human and primate infections (4, 53–56). Rigorous studies comparing the disease in animal models to that of human disease are crucial to defining the validity of each model system.

The second critical shortcoming in our portfolio is the complete absence of any biomarkers for disease status (57). In mice, one can isolate tissues and count colony-forming units to determine the progression of disease in the context of bacterial replication, but we cannot do this in

humans. There is no simple, noninvasive means of assaying whether an individual is containing his or her infection or progressing to active disease. The diagnostic tests in common usage today merely measure whether or not an individual has mounted an immune response to *Mtb*. Given that progression to disease is determined locally, at the level of the individual granuloma, it is unclear if we will ever be able to develop biomarkers for disease progression based on systemic readouts of immune status (58). Recent advances in positron emission tomography (PET)– and computerized axial tomography (CAT)–based imaging modalities have the potential to fill this gap (Fig. 3), but the relation between the observed pathology and bacterial burden has yet to be established. Because such imaging is not a practical method for diagnosing TB in the field, studies correlating comparable imaging data with immune responses or metabolites measurable in readily obtainable human samples (blood and/or urine) should be undertaken in an attempt to identify possible biomarkers of infection and disease. This is an extremely serious gap in moving new therapies into the field where, currently, impact becomes apparent only after many years.

Conclusions and the Cold Slap of Reality

The development of new drugs and vaccines is sorely needed, but they are not the complete answer. Tuberculosis is an infection that can be

held in check by an effective public health system. Failure of the local health-care system can lead to the systematic selection and spread of drug-resistant strains, as recently observed in the KwaZulu-Natal province of South Africa, where extensively drug-resistant (XDR) strains resistant to as many as seven drugs are being observed (59–61). During the period leading up to the identification of these XDR strains, the health-care program in this region was in disarray, with only 18% of sputum-positive and 29% of hospitalized TB patients completing their treatment regimen. The coincidence of HIV was 80% in TB patients in this region, and there was no testing for drug-resistance conducted on the *Mtb* strains from patients. It is postulated that failed treatment regimens, in combination with the absence of drug-resistance testing, led to the selection and spread of these lethal strains.

The importance of an appropriately resourced, effectively managed public health system cannot be overstated. Even the best drugs or vaccines in the world will fail without an effective health-care infrastructure to administer and monitor their application. It is only through the combination of drugs, vaccines, and public health surveillance that we can hope to break the cycle of transmission, whereby each TB patient is responsible for one or more new cases of active disease.

References and Notes

1. J. L. Flynn, J. Chan, *Trends Microbiol.* **13**, 98 (2005).
2. D. G. Russell, *Nat. Rev. Microbiol.* **5**, 39 (2007).
3. D. G. Russell, P. J. Cardona, M. J. Kim, S. Allain, F. Altare, *Nat. Immunol.* **10**, 943 (2009).
4. L. E. Via *et al.*, *Infect. Immun.* **76**, 2333 (2008).
5. G. Kaplan *et al.*, *Infect. Immun.* **71**, 7099 (2003).
6. S. Y. Eum *et al.*, *Chest* **137**, 122 (2010).
7. C. E. Barry 3rd *et al.*, *Nat. Rev. Microbiol.* **7**, 845 (2009).
8. J. Dietrich, T. M. Doherty, *APMIS* **117**, 440 (2009).
9. J. D. MacMicking *et al.*, *Cell* **81**, 641 (1995).
10. C. Nathan, *Science* **312**, 1874 (2006).
11. L. G. Bekker *et al.*, *Infect. Immun.* **68**, 6954 (2000).
12. P. L. Lin *et al.*, *Arthritis Rheum.* **62**, 340 (2010).
13. H. L. Plessner *et al.*, *J. Infect. Dis.* **195**, 1643 (2007).
14. L. Tsenova, A. Bergtold, V. H. Freedman, R. A. Young, G. Kaplan, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 5657 (1999).
15. A. Fortin, L. Abel, J. L. Casanova, P. Gros, *Annu. Rev. Genomics Hum. Genet.* **8**, 163 (2007).
16. G. F. Black *et al.*, *Lancet* **359**, 1393 (2002).
17. I. Verma, A. Grover, *Expert Rev. Vaccines* **8**, 1547 (2009).
18. M. A. Behr, P. M. Small, *Nature* **389**, 133 (1997).
19. L. Brandt *et al.*, *Infect. Immun.* **70**, 672 (2002).
20. E. Lozes *et al.*, *Scand. J. Immunol.* **46**, 16 (1997).
21. C. Demangel, T. Garnier, I. Rosenkrands, S. T. Cole, *Infect. Immun.* **73**, 2190 (2005).
22. M. A. Horwitz, *Microbes Infect.* **7**, 947 (2005).
23. A. Williams *et al.*, *Tuberculosis (Edinb.)* **85**, 29 (2005).
24. L. Grode *et al.*, *J. Clin. Invest.* **115**, 2472 (2005).
25. D. Aguilar *et al.*, *Clin. Exp. Immunol.* **147**, 330 (2007).
26. J. Hinchey *et al.*, *J. Clin. Invest.* **117**, 2279 (2007).
27. W. R. Waters *et al.*, *Vaccine* **25**, 7832 (2007).
28. T. Elvang *et al.*, *PLoS ONE* **4**, e5139 (2009).
29. Y. A. Skeiky *et al.*, *Vaccine* **28**, 1084 (2010).
30. E. Z. Tchilian *et al.*, *Infect. Immun.* **77**, 622 (2009).
31. H. M. Vordermeier *et al.*, *Infect. Immun.* **77**, 3364 (2009).
32. K. T. Whelan *et al.*, *PLoS ONE* **4**, e5934 (2009).

33. S. G. Reed *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 2301 (2009).
34. A. L. de Noronha, A. Báfica, L. Nogueira, A. Barral, M. Barral-Netto, *Pathol. Res. Pract.* **204**, 155 (2008).
35. M. J. Doenhoff, *Immunol. Today* **19**, 462 (1998).
36. E. Ledru, S. Ledru, A. Zoubga, *Immunol. Today* **20**, 336 (1999).
37. I. Parwati, R. van Crevel, D. van Soolingen, *Lancet Infect. Dis.* **10**, 103 (2010).
38. I. C. Shampura *et al.*, *J. Clin. Microbiol.* **48**, 387 (2010).
39. E. J. Muñoz-Elías *et al.*, *Infect. Immun.* **73**, 546 (2005).
40. R. J. Rees, P. D. Hart, *Br. J. Exp. Pathol.* **42**, 83 (1961).
41. W. P. Gill *et al.*, *Nat. Med.* **15**, 211 (2009).
42. C. M. Sasseti, D. H. Boyd, E. J. Rubin, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 12712 (2001).
43. C. M. Sasseti, E. J. Rubin, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 12989 (2003).
44. J. D. McKinney *et al.*, *Nature* **406**, 735 (2000).
45. E. J. Muñoz-Elías, A. M. Upton, J. Cherian, J. D. McKinney, *Mol. Microbiol.* **60**, 1109 (2006).
46. A. K. Pandey, C. M. Sasseti, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 4376 (2008).
47. A. M. Upton, J. D. McKinney, *Microbiology* **153**, 3973 (2007).
48. R. Van der Geize *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 1947 (2007).
49. X. Yang, N. M. Nesbitt, E. Dubnau, I. Smith, N. S. Sampson, *Biochemistry* **48**, 3819 (2009).
50. P. Peyron *et al.*, *PLoS Pathog.* **4**, e1000204 (2008).
51. T. R. Rustad, A. M. Sherid, K. J. Minch, D. R. Sherman, *Cell. Microbiol.* **11**, 1151 (2009).
52. T. Christophe *et al.*, *PLoS Pathog.* **5**, e1000645 (2009).
53. P. L. Lin *et al.*, *Infect. Immun.* **74**, 3790 (2006).
54. P. L. Lin *et al.*, *Infect. Immun.* **77**, 4631 (2009).
55. L. H. Ly, D. N. McMurray, *Indian J. Exp. Biol.* **47**, 432 (2009).
56. Y. C. Manabe *et al.*, *Tuberculosis (Edinb.)* **88**, 187 (2008).
57. M. Doherty, R. S. Wallis, A. Zumla; WHO-Tropical Disease Research/European Commission joint expert consultation group, *Curr. Opin. Pulm. Med.* **15**, 181 (2009).
58. T. M. Doherty, R. S. Wallis, A. Zumla, *Clin. Chest Med.* **30**, 783 (2009).
59. M. Loveday, L. Thomson, M. Chopra, Z. Ndlela, *Int. J. Tuberc. Lung Dis.* **12**, 1042 (2008).
60. K. D. Jones, T. Hesketh, J. Yudkin, *Trans. R. Soc. Trop. Med. Hyg.* **102**, 219 (2008).
61. M. Pillay, A. W. Sturm, *Clin. Infect. Dis.* **45**, 1409 (2007).
62. Research in the authors' laboratories is supported by NIH grants AI067027, AI057086, AI080651, HL055936, and HL100928 (D.G.R.) and AI50732, HL71241, HL092883, and HL075845 (J.L.F.) and by the Intramural Research Program of the NIAID, NIH (C.E.B.). All three authors receive support from the Bill and Melinda Gates Foundation.

10.1126/science.1184784

REVIEW

The Population Dynamics and Control of Tuberculosis

Christopher Dye^{1*} and Brian G. Williams²

More than 36 million patients have been successfully treated via the World Health Organization's strategy for tuberculosis (TB) control since 1995. Despite predictions of a decline in global incidence, the number of new cases continues to grow, approaching 10 million in 2010. Here we review the changing relationship between the causative agent, *Mycobacterium tuberculosis*, and its human host and examine a range of factors that could explain the persistence of TB. Although there are ways to reduce susceptibility to infection and disease, and a high-efficacy vaccine would boost TB prevention, early diagnosis and drug treatment to interrupt transmission remain the top priorities for control. Whatever the technology used, success depends critically on the social, institutional, and epidemiological context in which it is applied.

A century or more of social and economic progress, reinforced by the postwar discovery of efficacious drugs, had driven tuberculosis (TB) to low levels in the rich world of the 1980s. With no systematic evaluation of data from developing countries, no forewarning of the impending spread of HIV/AIDS or of the demise of the Soviet Union, and no coherent approach to control, TB was invisible to international donors and taken to be a fact of life in the most-affected parts of the world.

Four events put TB firmly on the global health agenda. First, the 1990 Global Burden of Disease Study (GBD 1990) identified TB as one of the top 10 causes of morbidity and mortality worldwide. TB became prominent in health statistics because untreated disease, with a case fatality rate of around 50%, cost millions of young adults decades of healthy life. Second, GBD 1990 and

subsequent analyses began to measure the impact of the spread of HIV/AIDS and the health consequences of social and economic crisis in eastern Europe. Third, clinical and economic studies showed that combination drug treatment for TB was among the most effective and cost-effective of all health interventions. Fourth, in response to all these observations, the World Health Organization (WHO) launched a new control strategy, based on Directly Observed Treatment and Short-course drug therapy (DOTS).

In 2006 the Stop TB Strategy added new elements to DOTS, articulating the importance of managing drug-resistant and HIV-associated TB and of engaging the many different participants in health care. Twenty years on, systematic monitoring and evaluation have revealed both successes and failures of DOTS and the Stop TB Strategy. Among the successes, 36 million patients were treated worldwide between 1995 and 2008, and up to 8 million deaths were averted (*1*). WHO's target cure rate of 85% was exceeded in the 2007–2008 global cohort of 2.7 million new sputum smear-positive patients, and case detection in 2008 reached an estimated 61%, close to the 70%

target. A combination of surveys, surveillance, and mathematical modeling suggests that targets of halving 1990 levels of prevalence and mortality by 2015 could be reached in four of six WHO regions. The exceptions are Africa (sub-Saharan) and Europe (former Soviet countries).

Worldwide, TB incidence per capita is falling at an estimated 1% per year. Although this slow rate of decline satisfies the Millennium Development Goal (MDG) for TB, which is to ensure that the incidence rate is falling by 2015, the world's population is growing at about 2% per year, so the total number of new TB cases is still rising (*1*). There are expected to be 9.8 million new cases in 2010, more than in any previous year in history. Eighty percent of these cases will be found in the 20 to 25 highest-burden countries, and more than one-third in India and China. A review of cases reported by 134 countries between 1998 and 2007 found that only 35 had per capita rates of decline exceeding 5% per year (*2*). These were countries with either small populations (<20 million) or high incomes [gross domestic product (GDP) > US\$10,000 per capita]. Incidence rates were increasing in 41 countries, 19 in sub-Saharan Africa.

With MDG target year 2015 now in sight, we examine here a range of factors that affect TB burden and trends, and which could explain why drug treatment programs have had such little impact on TB transmission and case load.

TB Epidemiology and Control: The Standard Model

The expected effects of control are derived from the standard model of TB natural history in which the pathogen, *Mycobacterium tuberculosis*, is considered a single entity and the response to lung infection is represented by two dichotomies: fast (primary progression) or slow transition (via latency) from infection to infectious or noninfectious disease (Fig. 1) (*3*). This compartmentalization simplifies the natural history, makes epidemiological calculations easier, and facilitates the management of patients who can be placed in a few distinct categories. DOTS and the Stop TB

¹World Health Organization, CH1211 Geneva 27, Switzerland.

²South African Centre for Epidemiological Modelling and Analysis (SACEMA), Stellenbosch 7600, South Africa.

*To whom correspondence should be addressed. E-mail: dyec@who.int

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

The following resources related to this article are available online at www.sciencemag.org (this information is current as of May 13, 2010):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:
<http://www.sciencemag.org/cgi/content/full/328/5980/856>

Supporting Online Material can be found at:
<http://www.sciencemag.org/cgi/content/full/328/5980/856/DC1>

This article **cites 45 articles**, 12 of which can be accessed for free:
<http://www.sciencemag.org/cgi/content/full/328/5980/856#otherarticles>

This article appears in the following **subject collections**:
Medicine, Diseases
<http://www.sciencemag.org/cgi/collection/medicine>
Microbiology
<http://www.sciencemag.org/cgi/collection/microbio>

Strategy have given priority to patients that present at clinics with infectious, sputum smear-positive TB, as reflected in the WHO targets for case detection and cure.

The standard model has been formalized mathematically in various ways (4); one interpretation predicts that TB incidence per capita will fall at 5 to 10% per year when target values of case detection (70%) and cure (85%) have been exceeded (5). These results were derived mainly from the large-scale impact of drug treatment in Europe and North America after 1950. In England and Wales, as in the United States, the number of new TB cases per capita fell by 5 to 6% annually for 30 years and TB deaths at 9 to 10% annually (Fig. 2A). On a smaller scale, among Eskimo communities, intensive drug treatment supplemented by isoniazid preventive therapy has achieved even faster rates of decline in cases (13% per year) and deaths (30% per year) (6).

These early achievements in Europe and North America defined expectations for high-burden countries in the developing world, and there have been a few successes. To pick just one, Estonia has reversed the post-Soviet rise in TB, cutting the numbers of cases and deaths and case fatality, and pushing down multidrug-resistant TB (MDR-TB) at least as quickly as drug-sensitive TB (Fig. 2B) (7, 8). Estonia has shown what can be done with a good control program, but their achievement remains exceptional. Nearly two decades after the formulation of DOTS, and more than half a century after the development of combination therapy, the potential of drugs to reduce transmission has been only partly realized (Fig. 2, C and D) (1). We now look for explanations, by examining characteristics of the pathogen and its transmission, of humans as hosts, and of the structure of pathogen and human populations.

Certain Strains of *M. tuberculosis* Are More Transmissible Than Others

Before phenotypic (e.g., sensitivity to drugs) and genotypic (e.g., DNA fingerprinting) variation among strains could be observed, there was some justification for treating *M. tuberculosis* as a single organism. But as the analysis of genetic variation has become faster and cheaper, using a greater range of discerning techniques, much information has emerged about *M. tuberculosis* as a family of strains and lineages. Some of this is old biology, newly uncovered; other findings show how TB epidemiology is changing in the world today. Recent investigations have revealed hypervariable regions of the genome (9), the evolutionary origins of the *M. tuberculosis* complex (10), how this pathogen has spread around the world from African origins (11), and which genes code for drug resistance (12, 13). The data have demonstrated variation in the evolutionary fitness of drug-resistant Beijing and other strains, that different strains spread more or less

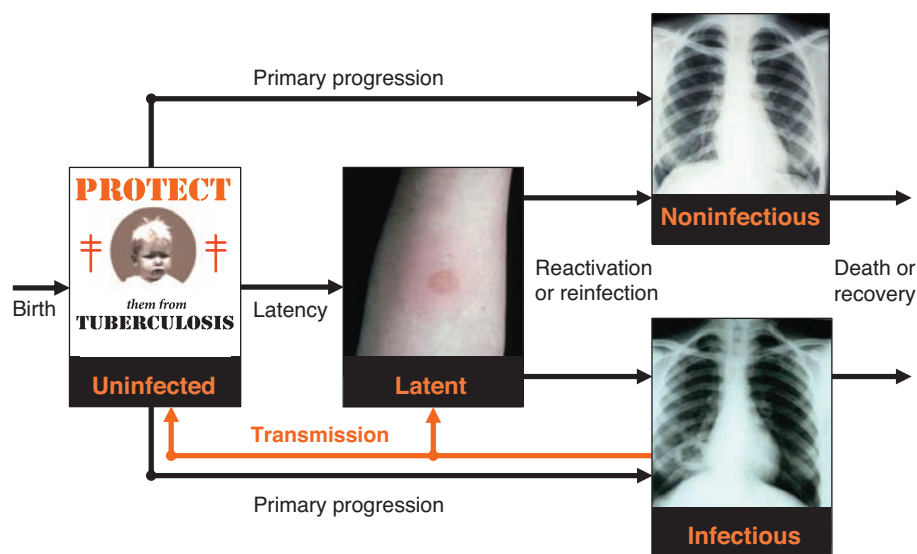


Fig. 1. Compartmental model of tuberculosis transmission, infection, reinfection, reactivation, and disease. Latent infection is commonly identified by a tuberculin skin test. Cavitory lung disease is most infectious (lower chest x-ray); noncavitory (upper chest x-ray) and extrapulmonary disease are less infectious or noninfectious. In the simplest mathematical models, TB is assumed to exist in a closed population, with no immigration or emigration and a constant number of people at each age (3). [Credits: "Protect" poster: American Lung Association/National Library of Medicine; "Latent": Bart's MedicalLibrary/Phototake; x-rays: Wellcome Images]

quickly through populations, and that there are strains more and less capable of causing active TB (7, 14–18) (Fig. 3A). From all this work it is clear that parameters of the standard model, as they relate to the characteristics of *M. tuberculosis*, vary among strains and settings, and through time. The magnitude of this variation and its epidemiological consequences are largely unquantified, though the success of control programs may depend on it.

Numerous Human Genetic Polymorphisms Are Linked to TB

High death rates from TB in 17th- to 19th-century Europe (≈ 20 to 30% of all mortality) exerted relatively strong selection pressure for genes that protect against TB, as compared with other infections. Even so, it is unlikely that selection was strong enough to explain much of the decline in TB in 20th-century Europe (Fig. 2A) (19). Now, in the early 21st century, mortality rates are far lower than in 19th-century Europe, even in countries considered to be highly endemic for TB (e.g., $\approx 3\%$ of deaths are due to TB in Southeast Asia). We can infer that natural selection has a negligible role in the decline of TB today.

More plausibly, geographically varying rates of selection in the past have influenced the present distribution of polymorphisms linked to TB. A variety of genetic variants are known to affect susceptibility to TB, including those of the major histocompatibility complex, the macrophage protein SLC11A1 (also known as NRAMPT), the vitamin D receptor, and the interferon- γ and nitric oxide synthase pathways (20). Case control and genome-wide association studies will doubtless

reveal more genetic mechanisms of resistance, but effects of modest size (odds ratios, OR < 3) will probably be the norm. Even with weak effects, common genes can account for a large fraction of cases in some settings. For example, children carrying the NRAMPT NO2 polymorphism at one site in the southern United States were three times as likely to have TB. Because of its high prevalence (76%) in that population, this variant accounted for most (85%) childhood TB (21). The main motive for studying human genetic determinants of TB has been to devise better tools for control, including vaccines and personalized drug regimens. Consequently, there have been few studies of how polymorphisms can explain contemporary epidemiological differences among populations—the background to control in any setting.

From an evolutionary perspective, we expect the outcome of infection to be influenced by gene-gene interactions within hosts and by interactions between pathogen and host genotypes. New evidence that *M. tuberculosis* lineages are adapted to specific host populations is therefore consistent with expectations (20, 22) (Fig. 3B). Given multiple polymorphisms and gene interactions, whose expression depends on environmental context, we anticipate a greater variety of responses to infection than are captured by the dichotomies of the standard model (Fig. 1).

Some Infected People Are at High Risk of Developing Active TB

The tuberculin skin test and interferon- γ release assays are measures of cellular immunity and

Tuberculosis & Malaria

putative markers of latent infection (23). These tests monitor a diversity of responses to infection, with varying rates of progression and recession including states with no persistent infection but primed T cells; low populations of nonreplicating bacteria; and replicating bacteria kept at sub-clinical levels by immunity (24). Individuals that have been exposed to *M. tuberculosis*, with various environmental and genetic backgrounds, might pass through any number of these states, and the variable (among individuals) and changing (through time) response to infection could be an attribute of pathogen or host, or of the interplay between combinations of hosts and pathogens (Fig. 3B) (25). This unfolding story of latency is another reminder that the fast-slow dichotomy in Fig. 1 is a coarse representation of a continuum. And the balance of different kinds of latent state might vary from one population to another depending on, for example, the ambient and recent history of infection.

Isoniazid preventive therapy (IPT) is cheap and highly efficacious (up to 90%) in stopping progression to active TB among people who are skin-test positive. IPT is recommended for HIV-positives in TB endemic areas, though few patients receive it (26). It is not widely used among HIV-negatives because hundreds of infected people have to be tuberculin tested and then treated to prevent one case, the course of treatment is long (6 to 9 months), and there are side effects, albeit at low frequency. The result is that, despite having an effective drug, progression from latent to active TB is largely unchecked.

Some Kinds of People and Patients Transmit More Infections Than Others

The standard TB model distinguishes sputum smear-positive cases from smear-negative cases (Fig. 1). This routine differential diagnosis is not ideal because smear microscopy fails to confirm the presence of *M. tuberculosis* in many true TB cases (culture is more sensitive). But it does offer an underexploited opportunity to search actively for the most infectious cases in a community. In Morocco, for example, half the smear-positive cases are found among men aged 15 to 44 years, who make up only a quarter of the population and who are a focus of transmission, especially in densely populated urban areas (27).

The number of infections transmitted by a TB case depends not only on the bacterial load in sputum, but also on the duration of infectiousness. An episode of TB can be prolonged by events that take place before, during, and after diagnosis. There may be delays before diagnosis when health services are inaccessible, service charges are unaffordable, and procedures in infection control are not observed. Private, out-of-pocket expenditure is a major barrier to seeking health care and accounts for a large fraction of spending on health in poor countries (Fig. 3C). While traditional diagnostic tests have limited sensitivity

(sputum smear microscopy) and specificity (chest x-ray), further needless errors arise through their improper use. Some failures of standard procedure for preventing infection have had dramatic consequences, such as the 2008 South African outbreak of extensively drug-resistant TB (XDR-TB) caused mainly by hospital transmission (28).

More subtly, there could be an interaction between the severity of illness and a patient's behavioral response to it. If patients with moderate disease tolerate it for longer, they might transmit more infections during extended episodes of illness. The ratio of TB prevalence to incidence is a measure of the duration of illness, which increases markedly with age in some settings. The prevalence/incidence ratio exceeded 10 years among elderly people in the Republic of (South) Korea, which suggests that they were persistent sources of infection (29, 30). To the extent that disease severity is a characteristic of bacterial genotype, this is potentially a mechanism for the selection of strains of lower virulence.

Infectiousness does not always end with diagnosis. Although combination drug therapy has the potential to cure almost all patients carrying drug-sensitive and drug-resistant strains of *M. tuberculosis*, cutting transmission and preventing most TB deaths, some patients fail treatment because they do not complete the full course, or because they are given inappropriate combinations of drugs. The consequences can be serious, clinically and epidemiologically; for example, treatment failure is the most obvious explanation for the persistently high fraction ($\approx 10\%$) of patients with recurrent episodes of smear-positive TB in India (1).

Most HIV Epidemics Have Peaked, But Prevalence Is Falling Slowly

HIV coinfection is the most powerful known risk factor for susceptibility to *M. tuberculosis* infection and progression to active disease. The incidence rate ratio of TB among HIV-positives

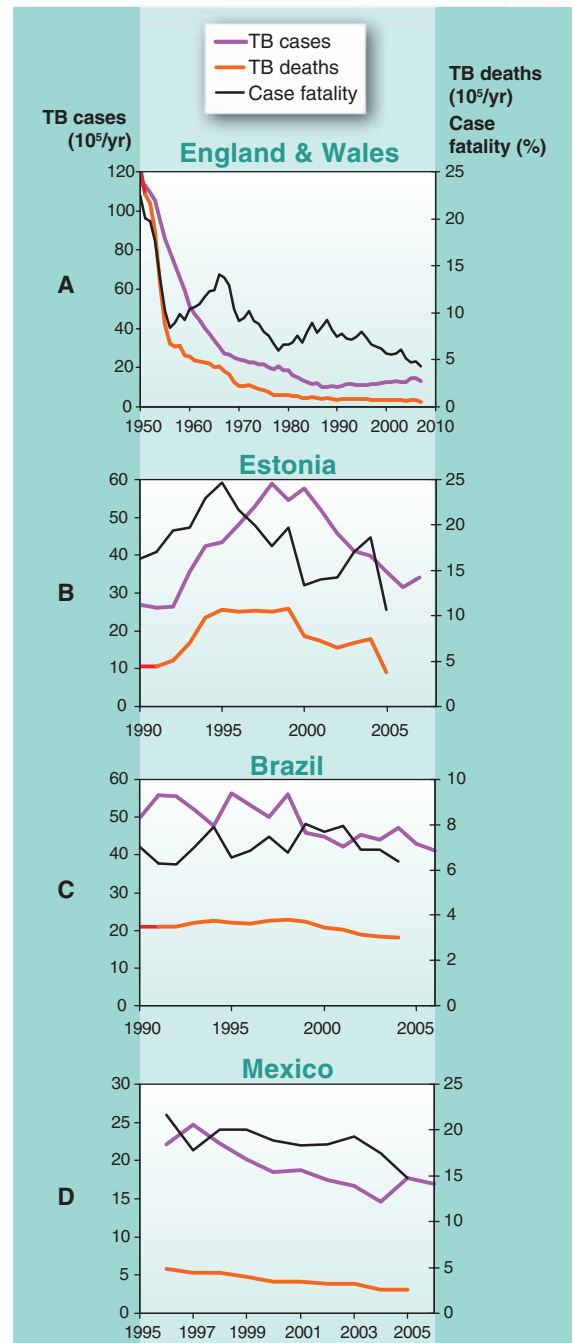


Fig. 2. Trends in TB cases and deaths per 100,000 population reported from (A) England and Wales (1913 to 2007), (B) Estonia (1990 to 2007), (C) Brazil (1990 to 2006), and (D) Mexico (1996 to 2006). The decline in European countries such as England and Wales (A) between 1950 and 1980 (cases 6%/year, deaths 10%/year) set expectations that have not been met in most developing countries. Estonia (B) is an unusual example of success in former Soviet Europe, showing how the rise of drug-sensitive and MDR-TB can be reversed by early diagnosis and treatment. Cases have fallen at 7%/year and deaths at 11%/year since the peak in 1998. TB decline has been sluggish in Brazil (C) (cases 2%/year, deaths 1%/year) and somewhat faster in Mexico (D) (cases 4%/year, deaths 7%/year), but still below potential. Data from the UK Health Protection Agency (www.hpa.org.uk) and (1, 7, 8).

as compared with HIV-negatives is on the order of 20 in countries with high HIV prevalence (1). The effect at population level has been to treble the number of TB cases reported annually from sub-Saharan Africa, and the increase has been greatest in eastern and southern parts of the continent. HIV epidemics have now peaked in the worst-affected African countries. In Tanzania, HIV seroprevalence data indicate that the HIV incidence rate was already falling by 1990, and prevalence by 1996. Consequently, the TB incidence rate has also been in slow decline since 2000 (Fig. 3D).

Although HIV epidemics have generated a large excess of TB cases, it is still not known whether coinfection has generated more or fewer *M. tuberculosis* infections than otherwise expected. Some studies have found that HIV-infected TB cases are less likely to be smear-positive than those cases uninfected with HIV, and that illness typically progresses more quickly, so that the infectious period before diagnosis is shorter. The “disease duration ratio” for HIV-positive TB cases as compared with HIV-negatives has been compared in two South African communities (31), but neither study calculated the total number of infections transmitted.

Epidemics of Some Chronic Diseases Exacerbate TB

Apart from HIV, other factors well known to increase susceptibility to infection and disease include alcohol abuse, diabetes, undernutrition, and tobacco smoking (32). The adverse effects of smoking are probably due to a combination of mechanical lung damage, oxidative stress, and the inhibition of T cell-mediated immune pathways that produce interferon- γ , interleukin-12, and tumor necrosis factor- α (33). Smokers typically have about twice the risk of pulmonary TB as compared with nonsmokers and, in India, up to half of all male TB deaths have been attributed to smoking (34). The prevalence of smoking in India, as in most other developing countries, is no longer increasing. But the number of smokers and smoking-related deaths is rising with population growth (35). Smoking is therefore expected to increase the number of TB cases and deaths too.

Like smoking, diabetes is associated with the suppression of cell-mediated immunity, though this might not fully explain the roughly threefold increase in TB risk among diabetics (36). Unlike smoking, both the prevalence of diabetes and the number of diabetes cases are rising worldwide. The increase is faster in developing than in high-income countries, with onset at younger ages in Asia (37, 38). Approximately 20% of smear-positive TB cases were attributed to concurrent diabetes in India in 2000 (39). If the projected rise from 25 million diabetes cases in 2000 to 80 million in 2030 is accurate, and the risk ratio remains the same, then 42% of smear-positive TB cases in India will be attributable to diabetes in 2030. Each TB case caused by diabetes would

infect other people, adding to the TB burden. Paradoxically, diabetes is linked to obesity, but whereas the former is a risk factor for TB, the latter is evidently protective (40). This apparent contradiction is yet to be resolved.

During the 1990s in post-Soviet Russia, alcoholism was a major cause of premature death and may also have contributed to the resurgence of TB (41). The number of overweight or obese people is increasing worldwide, but so is the number who are underweight. The number of undernourished people has grown by an estimated 9% since 1990, exceeding a billion in 2009, with unknown impact on TB (42).

Most People Now Live in High-Density Urban Areas

As often observed, infectious diseases are a penalty for high-density urban living; TB reached a peak in 19th-century Europe as urbanization went hand in hand with industrialization. In India, the risk of TB infection is higher in urban than in rural areas (43), and the urban population of India is expected to double between 2010 and 2030, reaching nearly 600 million people. But city life has advantages that offset the relatively high contact rates: Compared with rural areas, a higher proportion of people are wealthy and have better access to health services. In line with this view, the percentage of people in Indian cities living in poverty (as judged by income and living conditions) halved from 49% in 1973–1974 to 26% in 2004–2005 (44). Whether this downward trend in urban poverty will forestall an increase in TB remains an open question, not least because urbanization is confounded with other factors such as the rising prevalence of diabetes (38).

In China, internal migration has played a bigger part in urban growth than in India (45), with important consequences for TB. The number of TB cases per 100,000 inhabitants of Beijing declined from 20 in 1980 to 7 in 1993, but has since remained more or less steady as the number of cases among migrants to the city has increased.

In the low-incidence, high-income countries of western Europe and north America, immigration from high-burden countries has become a dominant factor in TB epidemiology. Most TB cases in the Netherlands, Norway, the United Kingdom, and the United States are due to infections brought in by young, foreign-born adults. These immigrants typically live in inner-city areas, often with poor access to health care. Consequently, TB incidence is falling more slowly in countries that have a larger proportion of foreign-born cases (2).

Populations Are Aging

Most of the countries where TB remains highly endemic are in demographic and epidemiological transition. Longevity is increasing, birth rates are falling, and populations are becoming older on average. The average age of the Indian population was 28 years in 2010 but will be 39 years in 2050.

Greater longevity is swelling the age classes that now have the highest TB incidence rates, namely adults over 45 years.

But it is not yet clear whether we can expect more or less TB in aging populations. With longer life spans, more people survive to acquire infection. With falling fertility, a population acquires fewer susceptibles each year. Greater longevity and lower fertility both tend to increase the prevalence of infection. Individuals carrying latent infections are a source of TB arising by reactivation, but they are also partially immune to reinfection. A simple elaboration of the standard TB model suggests that, on balance, aging will cause a small reduction in TB incidence per capita, unless the risk of developing TB after exposure is relatively high in older people (3). A relatively high risk of TB among the elderly is a real possibility, and could explain the slowing decline of TB in Hong Kong (46). To resolve this question, the standard model needs to be extended to explore the interplay between survival, fertility, and the risk of TB with age. Whether TB incidence (per capita or total case load) rises or falls in aging populations, the proportion of cases among elderly people is almost certain to increase in most countries.

Young Adults Are Rejuvenating TB Epidemics

The standard TB model assumes that patterns of mixing among infectious cases and their contacts, and the risk of TB among those infected, are constant through time. Consequently, when transmission is reduced by drug treatment, the average age of new cases should increase. The reason is that a progressively higher proportion of these cases comes from older rather than recent infections. In some countries where TB is in slow decline, like China, there are clear deviations from the expected pattern. The average age of older Chinese women (>55 years) with smear-positive TB has been steadily increasing, but the reverse is true for younger women with TB (Fig. 3E). Mexico, Myanmar, Sri Lanka, and Vietnam show the same phenomenon, for men or women or both.

There are at least four possible explanations. First, HIV coinfection generates a disproportionate number of cases among young men and women. Second, infections are imported into cities by young, foreign-born adults. Third, there are higher contact and transmission rates among young adults migrating internally to high-density urban areas. Fourth, novel strains like Beijing are found predominantly among younger cases and could have higher rates of transmission than other strains. Whatever the cause in any setting, the overrepresentation of TB in young adults is a departure from the standard model, and another obstacle to TB control.

Economic Shocks Can Cause Detrimental and Lasting Epidemiological Changes

Financial crises can interrupt health services while augmenting the risk of TB associated with, for ex-

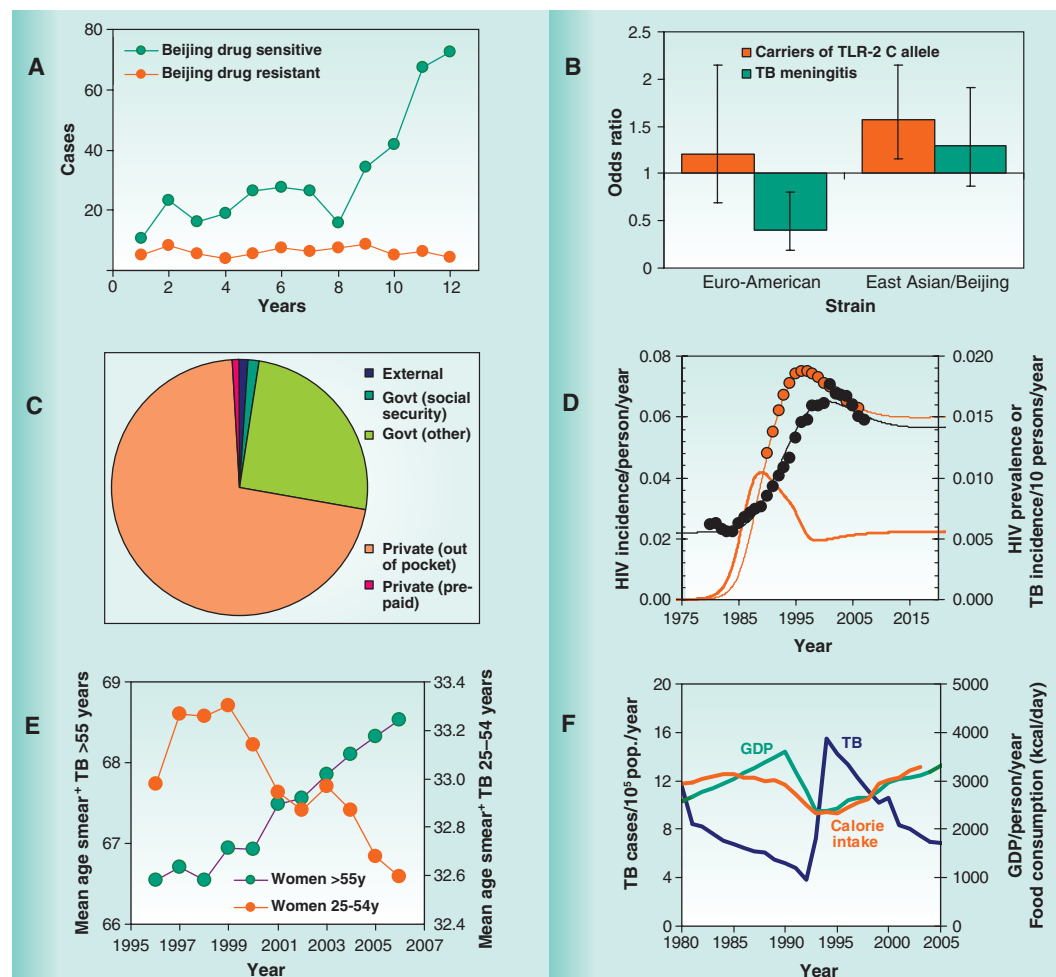


Fig. 3. Understanding the factors that affect TB trends will help to make control programs more successful. **(A)** Two strains of *M. tuberculosis* with different dynamics. Annual number of drug-sensitive (green) and drug-resistant (orange) cases of the Beijing clade. The number of cases caused by the drug-susceptible clade rose steeply after year 8. Data from (17). **(B)** Interactions between host and bacterial genotypes affect the risk of TB. In Vietnam, individuals carrying the C allele at the TLR-2 T597C locus appeared more likely to have TB caused by the East-Asian/Beijing genotype than by other genotypes (orange). Individuals infected with the Euro-American lineage of *M. tuberculosis* were less likely to have meningeal than pulmonary tuberculosis (green), suggesting that these strains are less capable of extrapulmonary dissemination within hosts. Data from (22). **(C)** Sources of health expenditure in India. Seventy-two percent of spending on health is private, out-of-pocket, a characteristic of many developing countries. For individuals on low incomes, an episode of TB can cost the equivalent of several months' salary. Source: WHO national health accounts (www.who.int/nha/en). **(D)** HIV prevalence has peaked in most countries, and TB epidemics are expected to turn downwards after a delay of 4 to 5 years. Here a mathematical model has been used to interpret and project trends in HIV incidence (orange line), HIV prevalence (orange line and points), and TB incidence (black line and points) in Tanzania. Source: authors' calculations. **(E)** Rejuvenation of the TB epidemic in China. As the risk of infection declines, the average age of smear-positive TB cases increases among elderly women (green), but falls among young women (orange). The disproportionately large number of cases among young adults is an obstacle to TB control. Data from (1). **(F)** Economic shock and tuberculosis in Cuba. After 1991, GDP (green) and food energy consumption (orange, kcal/day) dropped sharply, as TB incidence (blue) increased. TB cases started falling again slowly as GDP and calorie intake recovered (1, 47).

ample, alcoholism, malnutrition, and the congregation of displaced people. The impact on TB dynamics is strongly asymmetric because brief periods of elevated transmission, or of rapid breakdown from infection to disease, generate an abundance of new infections, which typically have an incubation period of several years. Reductions in incidence that took decades to achieve can be reversed within months, requiring decades for recovery. After the dissolution of the Soviet Union in

1991, excess TB cases and deaths in the newly independent states were tightly linked to lost economic productivity, and countries with close Soviet ties were also vulnerable, notably Cuba (Fig. 3F) (47). During the early 1990s in Cuba, protein, fat, and calorie intake fell with GDP, and there was a rise in the proportion of children born underweight. Almost 20 years later, TB incidence in the former Soviet countries and Cuba has still not returned to the low levels reported in 1991.

have higher reproductive fitness. To account for these factors will require better data, and structural and quantitative adjustments to the standard model in Fig. 1.

For 30 years, HIV coinfection has been the most prominent cause of increasing susceptibility to *M. tuberculosis* infection, and of rapid progression from infection to disease. Other aggravating factors are the rising numbers of alcoholics, diabetics, smokers, and elderly and undernourished

These observations leave a series of questions unanswered. First, the relative contribution of failing health services and failing health to excess morbidity and mortality is not known. Second, the characteristics of countries and populations that are vulnerable to these extreme events—for example, the nature of their health care systems—have not been defined. Third, these extreme events might point to smaller but perhaps more widespread social, economic, and medical determinants of TB trends, including chronic diseases.

TB Persistence and the Outlook for Control

In the face of widespread drug treatment, there are two possible reasons why the number of TB cases is still growing. Either transmission has not been reduced, or lower transmission rates have been offset by increasing susceptibility to infection and disease, augmented by population growth (Fig. 1).

There is plenty of evidence that transmission persists, from many settings. Delays before diagnosis can explain the propagation of infection in households, public spaces, clinics, and other points of congregation. Some of these delays are inherent to a system of control that has relied on patients to present themselves at health centers. And after diagnosis, treatment failure is a cause of long-lasting or recurrent episodes of infectious TB.

Although the traditional problems of case detection and case management remain, it is becoming clearer how and when social, economic, and evolutionary factors also affect transmission. One is the growth of cities, driven partly by internal and international migration. Another is the emergence of novel strains of *M. tuberculosis*, some of which

people. Large economic shocks, which affect both transmission and susceptibility, have shown how long-term gains in TB control can be swiftly and durably reversed. The gross effects of some economic and social crises have been evaluated retrospectively, but the timing and magnitude of such perturbations are practically impossible to anticipate.

Despite the uncertainties, we can draw some specific conclusions about the shorter- and longer-term prospects for TB control. The immediate priority is to pursue the goals of DOTS and the Stop TB Strategy because, in the face of a rising case load and persistent transmission, the greatest clinical and epidemiological benefits will come from prompt diagnosis and effective treatment. The present instruments for achieving this are not ideal, but there is clearly potential to deploy them more effectively: to accurately diagnose drug-sensitive and drug-resistant cases, select efficacious drug regimens, prevent transmission in clinics and hospitals, and actively seek TB cases in high-risk populations.

Improved technology will support these efforts, led by new diagnostics (e.g., those based on nucleic acid amplification and antigen detection). A broader portfolio of drugs is also needed to contain resistance, along with shorter regimens (<6 months) to make it easier for patients to complete treatment. There are currently fewer than 10 novel drugs in clinical trials, but the widening pipeline is certain to provide some options for changing and shortening regimens.

As HIV epidemics continue to generate a large but slowly falling number of TB cases, the challenge for TB and HIV/AIDS control programs is to aggressively implement the methods that are known to work, including antiretroviral therapy (ART) and isoniazid preventive therapy (IPT) to prevent TB among HIV-positives, and active TB case finding among HIV-positives followed by appropriate treatment for both infections (26, 48). Widespread and frequent HIV testing, with immediate ART for those who test positive, has the potential to improve clinical outcomes and to reduce the number of both new HIV infections and TB cases (49). These are in addition to other methods for HIV prevention, including the use of condoms and male circumcision.

There are other options for reducing susceptibility to infection, and for slowing the progression to active disease, but they will be hard to implement quickly. If all smokers quit, there would be far fewer TB cases and deaths, especially in populous Asia (50). Given that the proportion of men who smoke has been falling at only 2% per year in the UK and the USA, more slowly in Japan, and hardly at all yet in China and India, the removal of smoking as a risk factor for TB is likely to take decades (51). The result is that, although there are major health benefits from tobacco control, diabetes prevention, and nutritional interventions, these can play only supporting roles in TB control.

New technology to remove or neutralize infection—a better vaccine than BCG (bacille Calmette-Guérin) or practical methods for treating latent infection—could radically alter the approach to TB control (52), shifting the emphasis from cure to prevention. But this change will not happen soon. A major objective of vaccine research in the next 5 years is to establish sites to carry out efficacy trials. Preventive therapy might be reinvigorated if it becomes possible to identify, with some immunological or genetic marker, a subset of people who have a high risk of progressing to active TB. The hard task is to design a test that can be used in frequent surveys; regular sampling is needed to detect individuals at high risk because they pass quickly through the latent state.

Whatever the social, economic, and epidemiological context, TB control programs rely on the health systems of which they are part. TB drugs, equipment, and specialists are essential, but so are the other features of health services: nursing staff, laboratory networks, policies on private and public health care, financial protection schemes, and methods to engage patients and civil society. The instruments of policy, as much as those of technology, provide ways to reverse the rise of TB.

In sum, we have given examples of persistent transmission and of enhanced susceptibility to infection and disease, but it is easier to find evidence of the former. We conclude that control programs have been less effective than expected in cutting transmission mainly because patients are not diagnosed and cured quickly enough. The priority now is not to abandon the basic principles of chemotherapy, but rather to implement them with greater vigor.

References and Notes

- World Health Organization, "Global Tuberculosis Control: A Short Update To The 2009 Report" (World Health Organization, Geneva, 2009).
- C. Dye, K. Lönnroth, E. Jaramillo, B. G. Williams, M. Raviglione, *Bull. World Health Organ.* **87**, 683 (2009).
- Further details of TB natural history and additional analytical details are available as supporting material on Science Online.
- N. Bacaër, R. Ouifki, C. Pretorius, R. Wood, B. Williams, *J. Math. Biol.* **57**, 557 (2008).
- C. Dye, G. P. Garnett, K. Sleeman, B. G. Williams, *Lancet* **352**, 1886 (1998).
- S. Grzybowski, K. Styblo, E. Dorken, *Tubercle* **57** (suppl.), S1 (1976).
- C. Dye, B. G. Williams, *Sci. Transl. Med.* **1**, 3ra8 (2009).
- World Health Organization, "MXDR-TB Surveillance and Response: 2010 Global Update" (World Health Organization, Geneva, 2010).
- C. R. McEvoy, P. D. van Helden, R. M. Warren, N. C. Gey van Pittius, *BMC Evol. Biol.* **9**, 237 (2009).
- N. H. Smith, R. G. Hewinson, K. Kremer, R. Brosch, S. V. Gordon, *Nat. Rev. Microbiol.* **7**, 537 (2009).
- S. Gagneux, P. M. Small, *Lancet Infect. Dis.* **7**, 328 (2007).
- M. H. Hazbón et al., *Antimicrob. Agents Chemother.* **50**, 2640 (2006).
- T. Pramanian et al., *Clin. Microbiol. Infect.* **14**, 446 (2008).
- T. Cohen, M. Murray, *Nat. Med.* **10**, 1117 (2004).
- S. Gagneux et al., *Science* **312**, 1944 (2006).
- B. C. de Jong et al., *J. Infect. Dis.* **198**, 1037 (2008).
- G. D. van der Spuy et al., *Tuberculosis (Edinb.)* **89**, 120 (2009).
- I. Parwati, R. van Crevel, D. van Soolingen, *Lancet Infect. Dis.* **10**, 103 (2010).
- M. Lipsitch, A. O. Sousa, *Genetics* **161**, 1599 (2002).
- M. Moller, E. de Wit, E. G. Hoal, *FEMS Immunol. Med. Microbiol.* **58**, 3 (2010).
- S. Malik et al., *Proc. Natl. Acad. Sci. U.S.A.* **102**, 12183 (2005).
- M. Caws et al., *PLoS Pathog.* **4**, e1000034 (2008).
- M. Pai, A. Zwerling, D. Menzies, *Ann. Intern. Med.* **149**, 177 (2008).
- C. E. Barry III et al., *Nat. Rev. Microbiol.* **7**, 845 (2009).
- M. J. Blaser, D. Kirschner, *Nature* **449**, 843 (2007).
- C. Akolo, I. Adetifa, S. Shepperd, J. Volmink, *Cochrane Database Syst. Rev.* **1**, CD000171 (2010).
- C. Dye, S. Ottmani, L. Laasri, N. Bencheikh, *Int. J. Tuberc. Lung Dis.* **11**, 1225 (2007).
- N. R. Gandhi et al., *Lancet* **368**, 1575 (2006).
- Y. P. Hong et al., *Tuber. Lung Dis.* **74**, 323 (1993).
- S. J. Kim, Y. P. Hong, W. J. Lew, S. C. Yang, E. G. Lee, *Tuber. Lung Dis.* **76**, 534 (1995).
- B. Williams, D. Maher, *Am. J. Respir. Crit. Care Med.* **175**, 6 (2007).
- M. Murray, in *Tuberculosis: The Essentials*, M. Raviglione, Ed. (Informa Healthcare, New York, 2009), pp. 23–59.
- M. Pai et al., *Expert Rev. Anti Infect. Ther.* **5**, 385 (2007).
- V. Gajalakshmi, R. Peto, T. S. Kanaka, P. Jha, *Lancet* **362**, 507 (2003).
- P. Jha et al., RGI-CGHR Investigators, *N. Engl. J. Med.* **358**, 1137 (2008).
- K. E. Dooley, R. E. Chaisson, *Lancet Infect. Dis.* **9**, 737 (2009).
- S. Wild, G. Roglic, A. Green, R. Sicree, H. King, *Diabetes Care* **27**, 1047 (2004).
- A. Ramachandran, R. C. Ma, C. Snehalatha, *Lancet* **375**, 408 (2010).
- C. R. Stevenson et al., *BMC Public Health* **7**, 234 (2007).
- K. Lönnroth, B. G. Williams, P. Cegielski, C. Dye, *Int. J. Epidemiol.* **39**, 149 (2010).
- D. Zaridze et al., *Lancet* **373**, 2201 (2009).
- C. B. Barrett, *Science* **327**, 825 (2010).
- V. K. Chadha, P. Kumar, P. S. Jagannatha, P. S. Vaidyanathan, K. P. Unnikrishnan, *Int. J. Tuberc. Lung Dis.* **9**, 116 (2005).
- V. M. Vashishtha, *Indian Pediatr.* **46**, 875 (2009).
- United Nations Population Fund, "State of World Population 2007: Unleashing the Potential of Urban Growth" (United Nations Population Fund, New York and Washington DC, 2007).
- E. Vynnycky, M. W. Borgdorff, C. C. Leung, C. M. Tam, P. E. Fine, *Epidemiol. Infect.* **136**, 943 (2008).
- A. Rodríguez-Ojea, S. Jiménez, A. Berdasco, M. Esquivel, *Public Health Nutr.* **5** (1A), 129 (2002).
- R. M. Granich, C. F. Gilks, C. Dye, K. M. De Cock, B. G. Williams, *Lancet* **373**, 48 (2009).
- K. Kranzer et al., *Lancet Infect. Dis.* **10**, 93 (2010).
- H. H. Lin, M. Murray, T. Cohen, C. Colijn, M. Ezzati, *Lancet* **372**, 1473 (2008).
- O. Shafey, M. Eriksen, H. Ross, J. Mackay, *The Tobacco Atlas* (World Health Organization, Geneva, ed. 3, 2010).
- L. J. Abu-Raddad et al., *Proc. Natl. Acad. Sci. U.S.A.* **106**, 13980 (2009).
- We thank P. Glaziou, M. Raviglione, and H. Rieder for helpful discussions. The authors alone are responsible for the views expressed in this paper, which do not necessarily represent the decisions, policy, or views of the World Health Organization.

Supporting Online Material

www.sciencemag.org/cgi/content/full/328/5980/856/DC1
SOM Text
Fig. S1
References
10.1126/science.1185449

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

The following resources related to this article are available online at www.sciencemag.org (this information is current as of May 13, 2010):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/328/5980/862>

This article **cites 38 articles**, 9 of which can be accessed for free:

<http://www.sciencemag.org/cgi/content/full/328/5980/862#otherarticles>

This article appears in the following **subject collections**:

Microbiology

<http://www.sciencemag.org/cgi/collection/microbio>

That Was Then But This Is Now: Malaria Research in the Time of an Eradication Agenda

Stefan H. I. Kappe,^{1,2*} Ashley M. Vaughan,¹ Justin A. Boddey,³ Alan F. Cowman³

The global research community must take up the challenge to work toward the eradication of malaria. In the past, malaria research has focused on drugs and vaccines that target the blood stage of infection, and mainly on the most deadly species, *Plasmodium falciparum*, all of which is justified by the need to prevent and treat the disease. This work remains critically important today. However, an increased research focus is now being placed on potential interventions that aim to kill the parasite stages transmitted to and by the mosquito vector because they may represent more vulnerable targets to stop the spread of malaria. Here, we highlight some of the research into malaria parasite biology that has the potential to provide new intervention targets for antimalarial drugs and vaccines.

More than 40% of the world's population lives with some risk of contracting malaria, with most recent estimates suggesting several hundred million clinical cases and 800,000 deaths each year. Malaria is caused mainly by four species of *Plasmodium* parasites, of which *Plasmodium falciparum* causes most disease and death across sub-Saharan Africa, and *Plasmodium vivax* is the most prevalent parasite in most other malaria-endemic parts of the world. Vivax malaria has been historically categorized as benign, but recent data indicate that severe disease caused by *P. vivax* infection is common (1). Current observations showing the increased transmission of nonhuman primate malaria parasites such as *Plasmodium knowlesi* to humans (2), the possibility that *P. falciparum* can propagate in great apes, and observations that apes harbour a diversity of parasite species closely allied with *P. falciparum* all raise concerns that animal reservoirs for human malaria will make control of human infection more difficult (3). Nevertheless, malaria appears to be retreating in some areas of the developing world, including Asia, South America, and even sub-Saharan Africa, where the human toll of malaria infection is highest and malaria transmission by *Anopheles gambiae* mosquitoes is efficient. This decrease in malaria is mainly attributable to expanded control programs using drug treatment of infected individuals, preventive drug treatment of populations at high risk of infection, and mosquito control with insecticide-treated bed nets and indoor-insecticide spraying.

¹Seattle Biomedical Research Institute, Seattle, WA 98109, USA. ²Department of Global Health, University of Washington, Seattle, WA 98195, USA. ³The Walter and Eliza Hall Institute of Medical Research, Melbourne 3050, Australia.

*To whom correspondence should be addressed. E-mail: stefan.kappe@seattlebiomed.org

dividual (the pre-erythrocytic stages), and when sexual-stage gametocytes are taken up by the mosquito from the blood of an infected individual, leading to colonization of the mosquito midgut (Fig. 1). The parasite numbers passing from mosquito to human and back to mosquito are exceedingly low (a few dozen to a few hundred) when compared with parasite populations in an infected individual during asexual blood-stage replication (billions) and thus may be easier to eliminate (Fig. 1). However, research on asexual blood stages of the parasite remains essential for the development of new therapeutics because it is this phase of the life cycle that causes disease and death. Here, we highlight recent progress in understanding the intricate biology of malaria parasites and how some of these findings may lead to new strategies for intervention.

New Drugs and Novel Targets

The erythrocytic cycle of *P. falciparum* involves a massive amplification of the parasite population through periodic cycles of invasion, growth, division, and egress from erythrocytes (Fig. 1). To develop new compounds that target the parasite blood stage, an increased understanding of the mode of action of current antimalarials is necessary (4). *Plasmodium* parasites catabolize erythrocyte hemoglobin in the food vacuole, releasing toxic heme, which is sequestered in the food vacuole as a crystalline structure called hemozoin

Table 1. Drugs and possible drug targets for antimalarials against blood-stage and liver-stage parasites. ND, life cycle stage not determined; BS, blood stage; LS, liver stage; DHFR, dihydrofolate reductase; DHPS, dihydropteroate synthase; SHMT, serine hydroxymethyltransferase.

Location/function	Drug	Life stage/target
Apicoplast	Ciprofloxacin	BS and LS*/DNA replication
	Clindamycin/doxycycline	BS and LS/protein translation
	Fosmidomycin	BS and LS*/isoprenoid biosynthesis
	Rifampicin	BS and LS*/RNA transcription
	Fatty acid synthesis inhibitors	LS/fatty acid biosynthesis
Mitochondria	Atovaquone	BS and LS/cytochrome bc ₁ complex
Parasite cytoplasm	Pyrimethamine/chlorproguanil	BS/DHFR
	Sulfadoxine	BS/DHPS
	None	BS/SHMT
Endoplasmic reticulum	None	BS and LS*/plasmepsin V
Food vacuole	Chloroquine/amodiaquine	BS/heme polymerization
	Mefloquine/quinine	BS/heme polymerization
	Artemisinin†	BS and gametocyte/heme polymerization‡
	None	BS/falcipains
	None	BS/plasmepsins
Parasitophorous vacuole	None	BS and LS*/putative translocon
Merozoite invasion	None	BS/rhomboid 4, subtilisin 2
Merozoite egress	None	BS and LS/DPAP3, SERA5, subtilisin 1
Unknown	Primaquine	LS and gametocytes/drug target not known

*No published data to confirm drug action. †Exact drug target is not known. ‡Possible drug target.

(Fig. 2). Antimalarials such as chloroquine appear to interfere with heme sequestration, and despite the rise of widespread resistance to chloroquine, heme sequestration remains a viable target for the development of new antimalarials (5). The important artemisinin group of antimalarials are thought to have their effect by activation of the endoperoxide group through the presence of free Fe^{3+} , which is released during digestion of hemoglobin, and new generations of compounds are being developed that target heme metabolism (5). Proteases are responsible for the degradation of hemoglobin, providing a source of amino acids for parasite protein synthesis as well as a supply of metabolic energy (Fig. 2 and Table 1). Plasmepsin protease inhibitors can potently block *P. falciparum* growth, but many proteases have redundant functions, suggesting that drugs directed to the food vacuole will need to inhibit multiple proteases (6).

Antifolate compounds, including pyrimethamine and sulfadoxine (Table 1), have long been an important antimalarial combination, although the rapid rise of drug resistance has greatly reduced their effectiveness (7). They inhibit enzymes involved in folate synthesis, which is essential for parasite viability (8), and the combination of pyrimethamine and sulfadoxine show synergism, increasing their effectiveness against *P. falciparum*. The crystal structure for *P. falciparum* dihydrofolate reductase/tetrahydrofolate synthetase (9), the target for pyrimethamine, has identified regions that are important for dimerization of this enzyme, and a charged groove is located on the surface of the dimer that appears to be important in channeling substrate between the two active sites of the enzyme's bifunctional domains. These features provide new strategies to develop novel antimalarials against this target. Additionally, analysis of the folate enzymatic pathway has identified serine hydroxymethyltransferase (SHMT) as a target for development of novel antimalarials that target the folate biosynthetic pathway (10).

Plasmodia actively invade and egress from erythrocytes, and these processes require specialized parasite proteins (Fig. 3). Inhibition of the action of these proteins provides a novel avenue for antimalarial development. The invasion process involves a cascade of ligand-receptor interactions that are required to recognize the appropriate host cell and establish a tight parasite-host con-

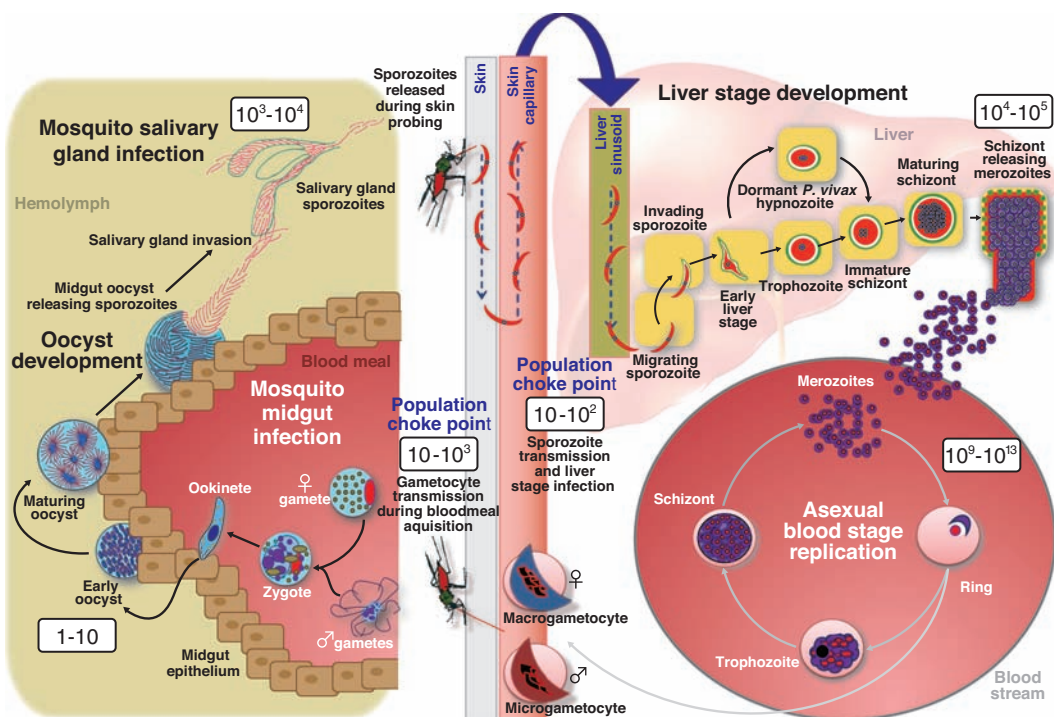


Fig. 1. The *Plasmodium* life cycle provides numerous potential points of intervention. Upon mosquito injection of motile salivary gland sporozoites into the human skin, parasites migrate in the tissue and invade a blood vessel. Sporozoites are transported to the liver via the blood stream, leave the liver sinusoid, and infect hepatocytes to initiate liver-stage infection. The parasite undergoes massive growth and forms tens of thousands of first-generation merozoites. Merozoites are released into the blood stream. Each merozoite infects an erythrocyte and initiates the intraerythrocytic cycle. The parasites replicate by means of cyclic infection, replication, and release of next-generation merozoites. Some merozoites form sexual-stage gametocytes after invasion, and these can be taken up by a mosquito during a blood meal. Gametes mate, form a zygote, and develop into an ookinete that infects the mosquito midgut. The resulting oocyst produces sporozoites, which invade the salivary glands, ready for transmission to the next host. Initial infection of the human host liver and the transmission of gametocytes to mosquitoes are considered choke points of the parasite life cycle because parasite numbers are lowest at these points. The numbers shown indicate parasite population sizes during life cycle progression.

tact that ultimately connects with the actomyosin motor inside the parasite to provide the force required for entry (11). Although the ability to invade host cells is an essential step, the parasite also has to escape, and this process of egress represents a potential target for drug development (12). The intraerythrocytic parasite is surrounded by the parasitophorous vacuole (PV) and the erythrocyte membrane. Both of these are ruptured during egress, allowing access to new erythrocytes for invasion. Egress occurs in a cascade of events in which the membranes rupture, resulting in the explosive release of merozoites (13, 14).

After invasion, the parasite begins a remarkable process of remodeling that converts the terminally differentiated erythrocyte, which lacks a nucleus and machinery for functions such as protein trafficking, into a niche in which the parasite obtains nutrients and hides from host protective mechanisms (15). This requires the export of hundreds of parasite proteins through the PV membrane (PVM) to the erythrocyte cytoplasm and, in some cases, to the host-cell membrane. As a result, large parasite-derived structures appear in the infected erythrocyte—for example,

Maurer's clefts, which are important for protein trafficking and sorting, as well as protrusions on the erythrocyte membrane called "knobs" (16). A pentameric motif is located at the N terminus of most exported proteins and is required for targeting beyond the PVM (15). This motif is a protease recognition sequence that is cleaved in the parasite's endoplasmic reticulum (ER) by an aspartic acid protease, plasmepsin V, revealing the protein's export signal (17, 18) for subsequent transport most likely through a translocon complex (19), making protein export a novel antimalarial target.

Can We Develop a Blood-Stage Vaccine?

Development of effective blood-stage vaccines is difficult because of antigenic diversity, parasite mechanisms that evade host responses, and the sheer biomass of parasites present within the host during an infection. Despite this, considerable efforts have been made to develop and test potential vaccine candidates from *P. falciparum* blood stages because evidence shows that repeated infection of humans results in control of blood-stage parasitaemia and effective immunity

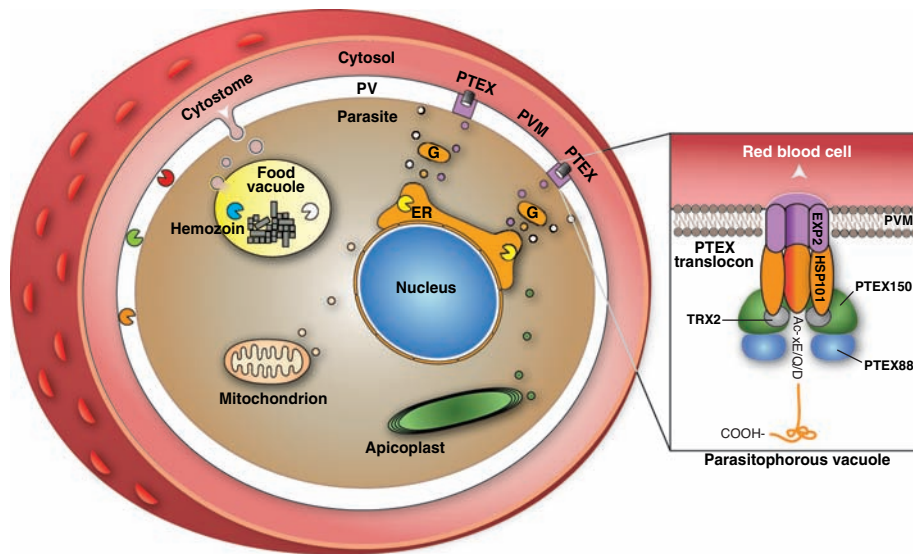


Fig. 2. Potential and realized drug targets of asexual *P. falciparum* blood stages. Shown are parasite organelles including the nucleus, apicoplast, mitochondrion, ER, Golgi (G), and food vacuole. The intraerythrocytic parasite is surrounded by the PV and PVM, through which proteins traffic to reach the erythrocyte. Most proteins for export are recognized by virtue of a PEXEL motif that is cleaved by Plasmepsin V in the ER to reveal the export signal. These proteins are carried to the PV, where they are recognized for export to the erythrocyte by a putative translocon (PTEX) in the PVM. The translocon consists of at least five subunits, including EXP2, HSP101, PTEX150, PTEX88, and TRX2 (inset). Egress of mature merozoites occurs by the breakdown of the PVM and erythrocyte membrane. This is mediated by proteases in the PV, including subtilisin 1, DPAP3, and SERA5 (solid partial circles of various colors). The food vacuole contains hemozoin, consisting of a crystalline form of heme, which results from the degradation of hemoglobin. Hemoglobin from the erythrocyte cytoplasm is endocytosed through the cytosome, and the endocytic vesicles fuse with the food vacuole membrane, releasing their contents into this structure for degradation by proteases such as plasmepsin and falcipains (solid partial circles).

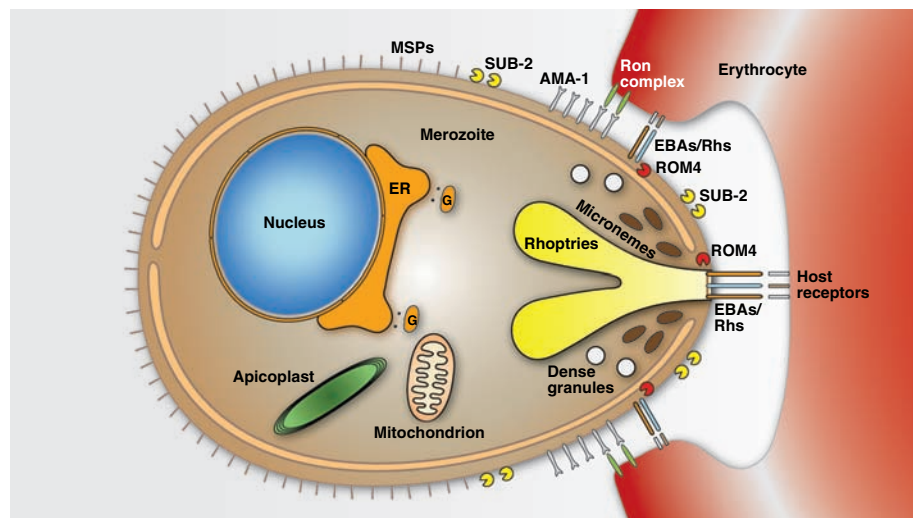


Fig. 3. *P. falciparum* merozoite invasion. The merozoite interacts with the erythrocyte at the apical end via parasite ligands (EBA-175, EBA-140, EBA-181, PfRh1, PfRh2a/b, PfRh4, and PfRh5) that bind to host receptors and activate the invasion process and formation of the tight junction. The tight junction binds the erythrocyte surface to the merozoite and moves across the surface of the invading parasite. The ligand-receptor interactions are released during invasion by cleavage of parasite ligands with rhomboid 4 (ROM4). Subtilisin 2 (SUB-2) is also involved in processing and shedding proteins on the merozoite surface during invasion. PfAMA-1 participates in the tight junction with the parasite RON complex that spans the erythrocyte membrane, tight junction, and merozoite membrane.

that reduces clinical disease (20). Antigens on the invasive merozoite (Fig. 3) have been considered prime vaccine candidates, but blood-stage vaccines have yet to show success (21). The stakes for blood-stage vaccines are even higher when malaria eradication is the aim because they must not only reduce disease but also parasite burden to a degree that reduces transmission.

Repeated low-dose infection of human volunteers with *P. falciparum* blood stages followed by drug treatment induces strong immunity against homologous challenge (22), and recently rodent malaria parasite blood stages were genetically attenuated through gene deletion of purine nucleoside phosphorylase (PNP) or nucleoside transporter 1 (NT1) (23, 24). These knockout strains protect against challenge with virulent parasites, and in the case of the NT1 knockout strain, complete protection was achieved against challenges with different parasite strains. Therefore, the approach of using genetically attenuated *P. falciparum* blood stages as potential vaccine warrants further research, and such vaccination also presents opportunities to identify correlates of protection.

Targeting the Apicoplast Organelle?

The apicoplast is unique to apicomplexan parasites, including *Plasmodium* spp. The bacteria-like origin as well as the essential functions contained within the apicoplast—including lipid, heme, and isoprenoid synthesis—represent a slew of potentially attractive drug targets (Fig. 2 and Table 1) (25). The apicoplast contains replication, transcription, and translation machineries, all of which represent targets for currently known inhibitors of these processes (Table 1). The apicoplast is maintained throughout the parasite life cycle; hence, the development of drugs that target its essential functions can potentially act against blood stages and liver stages, but this requires experimental target validation. For example, the utility of fatty-acid biosynthesis as a drug target for asexual blood stages has been recently brought into question because the enzymes in this pathway are not essential for this part of the life cycle (26–28). However, fatty-acid biosynthesis is critical for liver-stage development, suggesting that this pathway might be a drug target for pre-erythrocytic infection (26, 27).

Possible Targets for New Pre-Erythrocytic Antimalarials

Human malaria parasite transmission and liver infection is exceedingly difficult to study, and rodent malaria parasite models have proved invaluable for the study of pre-erythrocytic-stage infection. Infectious sporozoites leave the salivary glands of the mosquito during a bite and enter the host, make their way to the liver, invade hepatocytes, and commence development as liver stages (Fig. 1) (29). Sporozoites are injected into the avascular tissue of the skin then invade blood vessels (30),

and their extended stay in the skin leaves the sporozoite vulnerable to immune attack. Sporozoites traverse through cells by means of plasma membrane disruption, and they use this process to migrate through tissues and reach the liver (31). Once a sporozoite encounters a suitable hepatocyte in the liver, it invades, forms a PV, and initiates liver-stage development. Tissue migration and hepatocyte invasion is mediated by interactions of sporozoite proteins with host cell receptors, which might provide new avenues to block infection (29). In addition, rapid growth of liver stages depends on host nutrients that are actively taken from the infected hepatocyte. Hence, methods to starve the liver stage might be good avenues to prevent growth. Liver stages ultimately differentiate into tens of thousands of first-generation merozoites, which are released from the infected hepatocyte into the blood stream. The release is controlled by the parasite and occurs in packages that are still surrounded by the hepatocyte membrane (32). Egress from the liver is probably dependent on proteases, and indeed some of the same proteases necessary for blood-stage egress are expressed during late liver-stage development (33). Thus, protease inhibitors that target blood stages might also inhibit parasite release from the liver.

In essence, understanding host-parasite interactions during liver infection has the potential to inform new interventions that block this vulnerable phase of the parasite life cycle, and thus pre-erythrocytic-stage research should be given high priority.

Vaccines Against Pre-Erythrocytic Parasite Stages

A vaccine for the pre-erythrocytic stages is seen as the ideal tool for disease eradication. Furthermore, pre-erythrocytic stages appear not to exhibit substantial antigenic variation, and a vaccine based on a single parasite strain might provide protection from infection with heterologous strains. However, there is no solid evidence for naturally occurring protective immunity to pre-erythrocytic stages in malaria-endemic areas. Perhaps the repeated exposure to small numbers of sporozoites transmitted by each mosquito bite leads to a state of immune tolerance in the liver. Conversely, immunization with substantial numbers of irradiated sporozoites that are able to infect the liver but cannot complete liver-stage development elicits protective immunity that prevents subsequent infection, which has led to recent efforts in the manufacture of an irradiated *P. falciparum* sporozoite vaccine (34). Furthermore, genetically engineered sporozoites that are attenuated through the knockout of genes that are essential for liver-stage development provide a powerful new avenue for live attenuated parasite vaccine development. Such attenuated sporozoites confer complete protection in mouse models of malaria; recently, a genetically attenuated *P. falciparum* parasite with deletions in the

genes *P52* and *P36* was developed and will undergo clinical testing in the near future (35).

In parallel to whole-cell irradiated and genetically attenuated parasites, the design of effective subunit vaccines that target the sporozoite and liver stage are a high priority. The major immunodominant antigen of the sporozoite is the circumsporozoite protein (CSP), and a subunit vaccine based on CSP has proved partially successful in clinical trials (36). The major question now is whether one can identify additional protective pre-erythrocytic antigens and combine them with CSP so as to create a multi-subunit vaccine formulation that results in complete protection. The identification of novel targets for humoral protection is important, but even more critical is the identification of parasite targets that allow the immune system to eliminate infected hepatocytes through T cells (37). Novel pre-erythrocytic vaccine candidates can be tested for efficacy early in clinical development by using experimental malaria challenges of volunteers. Importantly, the experimental immunization with live-attenuated sporozoites provides a powerful platform to search for novel protective antigens by use of high-throughput immunological screens.

Transmission Blocking with a Vaccine?

Targeting the sexual stages of the malaria parasite life cycle with subunit vaccines could prevent transmission to the mosquito vector (38), and a greater focus on this life-cycle choke point is essential in an eradication agenda. Target antigens fall into two broad groups: those that are expressed before gamete fertilization and those that are expressed after fertilization on the surface of the ookinete—the invasive form that infects the mosquito midgut (Fig. 1). For example, immunization of nonhuman primates with the prefertilization antigen Pfs48/45 resulted in high antibody levels that block parasite transmission to mosquitoes in an ex vivo experiment that mixes the antisera with *P. falciparum*-infected blood and feeds the mixture to mosquitoes (39).

Unlike pre-erythrocytic- and erythrocytic-stage vaccine approaches, transmission-blocking vaccines do not protect the vaccinated individual, and some of the antigens are never seen naturally by the human immune system. Thus, the vaccine may not be boosted by malaria infection. As a result, vaccination must be potent and induce high antibody levels that are sustained for at least one transmission season. Moreover, initial testing of transmission-blocking vaccines for efficacy in clinical studies is not as straightforward as the testing of pre-erythrocytic vaccines because protection of the vaccinated individual is not the readout and population-based studies are needed.

Conclusions and Perspectives

Fundamental secrets of malaria parasite biology are now being revealed because of the development of genetic tools and the application of

sophisticated laboratory technologies. The recapitulation of complete parasite life cycles in the laboratory through use of rodent malaria models, the routine in vitro culture of *P. falciparum* blood stages, laboratory infection of mosquitoes with cultured gametocytes, and improved in vitro growth systems for liver stages has enabled research that was not previously possible. This, in combination with parasite whole-genome sequences, functional genomic tools, systems biology, and the ability to genetically manipulate the parasite, has led to an unprecedented wealth of data. Despite these advances, approximately 60% of *P. falciparum* genes encode proteins for which no functional assignments exist. Research on *P. vivax*, the major cause of malaria outside Africa, has lagged behind because the parasite is difficult to grow in the laboratory. *P. vivax* liver stages can lie dormant as “hypnozoites” and reactivate to cause a blood-stage infection months after the initial infection (Fig. 1), making the hypnozoite a high-priority target for research and drug development if malaria eradication is the goal. Currently, we do not have even the most basic understanding of the nature of hypnozoite dormancy. The search for new blood-stage malaria therapeutics continues, and knowledge of the pathophysiology and cell biology of this stage has vastly increased during the past decade, providing new intervention points to improve treatment of malaria.

However, the time has come to broaden our knowledge base of pre-erythrocytic and sexual stages in order to develop novel tools for blocking disease transmission. Promising research findings need to quickly move into human proof-of-concept studies for intervention. If proven, implementation must be accelerated, and this requires large investments. The intervention strategies discussed here need to be combined with other efforts that target the mosquito vector. Only then is it possible to envision a world without malaria within our lifetime.

References and Notes

1. R. N. Price et al., *Am. J. Trop. Med. Hyg.* **77** (suppl.), 79 (2007).
2. C. Putaporntip et al., *J. Infect. Dis.* **199**, 1143 (2009).
3. F. Prugnolle et al., *Proc. Natl. Acad. Sci. U.S.A.* **107**, 1458 (2010).
4. D. A. Fidock, R. T. Eastman, S. A. Ward, S. R. Meshnick, *Trends Parasitol.* **24**, 537 (2008).
5. T. N. Wells, P. L. Alonso, W. E. Gutteridge, *Nat. Rev. Drug Discov.* **8**, 879 (2009).
6. J. A. Bonilla, T. D. Bonilla, C. A. Yowell, H. Fujioka, J. B. Dame, *Mol. Microbiol.* **65**, 64 (2007).
7. Y. Yuthavong, S. Kamchonwongpaisan, U. Leartsakulpanich, P. Chitnumsub, *Future Microbiol.* **1**, 113 (2006).
8. T. Triglia, J. G. T. Menting, C. Wilson, A. F. Cowman, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 13944 (1997).
9. J. Yuwaniyama et al., *Nat. Struct. Biol.* **10**, 357 (2003).
10. N. Nirmalan, P. Wang, P. F. Sims, J. E. Hyde, *Mol. Microbiol.* **46**, 179 (2002).
11. A. F. Cowman, B. S. Crabb, *Cell* **124**, 755 (2006).
12. S. Yeoh et al., *Cell* **131**, 1072 (2007).
13. M. E. Wickham, J. G. Culvenor, A. F. Cowman, *J. Biol. Chem.* **278**, 37658 (2003).

14. S. Glushakova, D. Yin, T. Li, J. Zimmerberg, *Curr. Biol.* **15**, 1645 (2005).
15. M. Marti, J. Baum, M. Rug, L. Tilley, A. F. Cowman, *J. Cell Biol.* **171**, 587 (2005).
16. B. S. Crabb *et al.*, *Cell* **89**, 287 (1997).
17. J. A. Boddey *et al.*, *Nature* **463**, 627 (2010).
18. I. Russo *et al.*, *Nature* **463**, 632 (2010).
19. T. F. de Koning-Ward *et al.*, *Nature* **459**, 945 (2009).
20. K. Marsh, S. Kinyanjui, *Parasite Immunol.* **28**, 51 (2006).
21. J. T. Bruder, E. Angov, K. J. Limbach, T. L. Richie, *Hum. Vaccin.* **6**, 54 (2010).
22. D. J. Pombo *et al.*, *Lancet* **360**, 610 (2002).
23. L. M. Ting, M. Gissot, A. Coppi, P. Sinnis, K. Kim, *Nat. Med.* **14**, 954 (2008).
24. A. S. Aly, M. J. Downie, C. B. Mamoun, S. H. Kappe, *Cell. Microbiol.* published online 20 January 2010 (10.1111/j.1462-5822.2010.01440.x).
25. S. A. Ralph *et al.*, *Nat. Rev. Microbiol.* **2**, 203 (2004).
26. A. M. Vaughan *et al.*, *Cell. Microbiol.* **11**, 506 (2009).
27. Y. Pei *et al.*, *Mol. Microbiol.* **75**, 957 (2010).
28. M. Yu *et al.*, *Cell Host Microbe* **4**, 567 (2008).
29. A. M. Vaughan, A. S. Aly, S. H. Kappe, *Cell Host Microbe* **4**, 209 (2008).
30. R. Amino *et al.*, *Nat. Med.* **12**, 220 (2006).
31. M. M. Mota *et al.*, *Science* **291**, 141 (2001).
32. A. Sturm *et al.*, *Science* **313**, 1287 (2006).
33. A. S. Tarun *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 305 (2008).
34. S. L. Hoffman *et al.*, *Hum. Vaccin.* **6**, 97 (2010).
35. K. M. VanBuskirk *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 13004 (2009).
36. P. Bejon *et al.*, *J. Med.* **359**, 2521 (2008).
37. R. Wang, J. D. Smith, S. H. Kappe, *Expert Rev. Mol. Med.* **11**, e39 (2009).
38. C. J. Sutherland, *Mol. Biochem. Parasitol.* **166**, 93 (2009).
39. D. R. Chowdhury, E. Angov, T. Kariuki, N. Kumar, C. J. Sutherland, *PLoS ONE* **4**, e6352 (2009).
40. We apologize to many in the field whose work we could not cite because of space constraints. S.H.I.K. is supported by a grant from the Foundation for the National Institutes of Health through the Grand Challenges in Global Health initiative and NIH grant R01 AI053709. A.F.C. is supported by the National Health and Medical Research Council (NHMRC) of Australia, NIH grant R01 A144008, and an Australia fellowship and is an International Scholar of the Howard Hughes Medical Institute. J.A.B. is supported by a Peter Doherty Postdoctoral Fellowship from NHMRC.

10.1126/science.1184785

REVIEW

The Selection Landscape of Malaria Parasites

M. J. Mackinnon^{1,2*} and K. Marsh^{1,2}

Malaria parasites have to survive and transmit within a highly selective and ever-changing host environment. Because immunity to malaria is nonsterilizing and builds up slowly through repeated infections, commonly the parasite invades a host that is immunologically and physiologically different from its previous host. During the course of infection, the parasite must also keep pace with changes in host immune responses and red-blood-cell physiology. Here, we describe the “selection landscape” of the most virulent of the human malaria parasites, *Plasmodium falciparum*, and the adaptive mechanisms it uses to navigate through that landscape. Taking a cost-benefit view of parasite fitness, we consider the evolutionary outcomes of the most important forces of selection operating on the parasite, namely immunity, host death, drugs, mosquito availability, and coinfection. Given the huge potential for malaria parasite evolution in the context of the recently renewed effort to eradicate malaria, a deeper understanding of *P. falciparum* adaptation is essential.

Plasmodium falciparum is the most virulent of the human malaria parasites, causes the death of more than one million children per year (1), and is currently the target of renewed and massive control campaigns. Past failures to eradicate malaria have been hampered by the evolution of drug resistance in the parasite and insecticide resistance in its mosquito vector (2). At the heart of the malaria control problem, however, is the parasite’s remarkable capacity for adaptation to its highly heterogeneous natural environment. During its two-host life cycle, *P. falciparum* undergoes 10 morphological transitions in five different host tissues, proliferates asexually within three of these, and must propagate sexually at each transfer between hosts.

Within each host, it is subject to strong selection by host defenses and undergoes severe population bottlenecks. The parasite can infect most of its host population by reinventing people who have already mounted immune responses to it during previous or existing infections, and from each infection it can transmit for months, even years. Thus, this parasite has evolved the capacity to maximally exploit human beings for its own reproduction.

The recent explosion of genomics-based research has revealed just how uniquely adapted this parasite is to its hosts. Many of its genes have no known homologs in other species (3), and its machinery for interacting with the host environment is highly specialized (4). New sources of adaptive genetic and transcriptional variation in the parasite genome are beginning to be uncovered, and perhaps the most remarkable feature of the parasite is its adaptability per se. In this review, we explore the processes by which the parasite adapts in real time, with a view to understanding the potential impact control programs may have on parasite evolution.

Variation + Selection = Evolution

The framework we use to describe the process of adaptation in malaria parasites, summarized in Fig. 1, has the following components: (i) sources of variation (the substrate), (ii) selection (the mechanism), (iii) the fitness consequences (the outcome measure), (iv) scales on which adaptation operates (biological, temporal, and geographical), (v) constraints and trade-offs, and (vi) feedback between adaptation and the selective environment. To illustrate the multidimensional nature of the process, we use two examples, which are simplified to emphasize the contrasts.

The first is that of antigenic variation in the proteins displayed on the surface of the parasite-infected red blood cell (5). A major group of these variant surface antigens (VSAs) is coded for by a multigene family, *var*, containing ~60 copies of genes that are highly genetically diverse at the sequence level. The parasite expresses only one of these genes at a time, but each gene undergoes expression switching to alternative members. Thus, as the infection progresses, variant-specific antibodies to the expressed antigen develop, which afford a new parasite variant a competitive advantage. This variant rises in frequency until it, too, induces production of specific antibody against itself, thus causing diversifying selection. The fitness consequence of this immune avoidance is longer infections and, hence, greater opportunities to transmit to new hosts. The scale of selection is short-term and within-host, because the selected variant is not transmitted in the “on” state from the mosquito to the next host. There is rapid, direct feedback from the adapted parasite (the new variant) into its selective environment through the generation of variant-specific antibody. One possible constraint on this adaptive mechanism is a trade-off between antigenic novelty and the binding of parasitized red blood cells to host cells (cytoadherence), which plays a key role in the biology of *P. falciparum* (6, 7). Another potential constraint is a trade-off between the expression of high antigenic diversity within the lifetime of a single infection and reduced opportunities to colonize previously infected hosts carrying antibody to these types (8, 9).

¹Kenya Medical Research Institute–Wellcome Trust Research Programme, Post Office Box 230, Kilifi, Kenya. ²Centre for Tropical Medicine, Nuffield Department of Clinical Medicine, University of Oxford, Centre for Clinical Vaccinology and Tropical Medicine, Oxford OX3 7LJ, UK.

*To whom correspondence should be addressed. E-mail: mmackinnon@kilifi.kemri-wellcome.org

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

The following resources related to this article are available online at www.sciencemag.org (this information is current as of May 13, 2010):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/328/5980/866>

This article **cites 58 articles**, 17 of which can be accessed for free:

<http://www.sciencemag.org/cgi/content/full/328/5980/866#otherarticles>

This article appears in the following **subject collections**:

Microbiology

<http://www.sciencemag.org/cgi/collection/microbio>

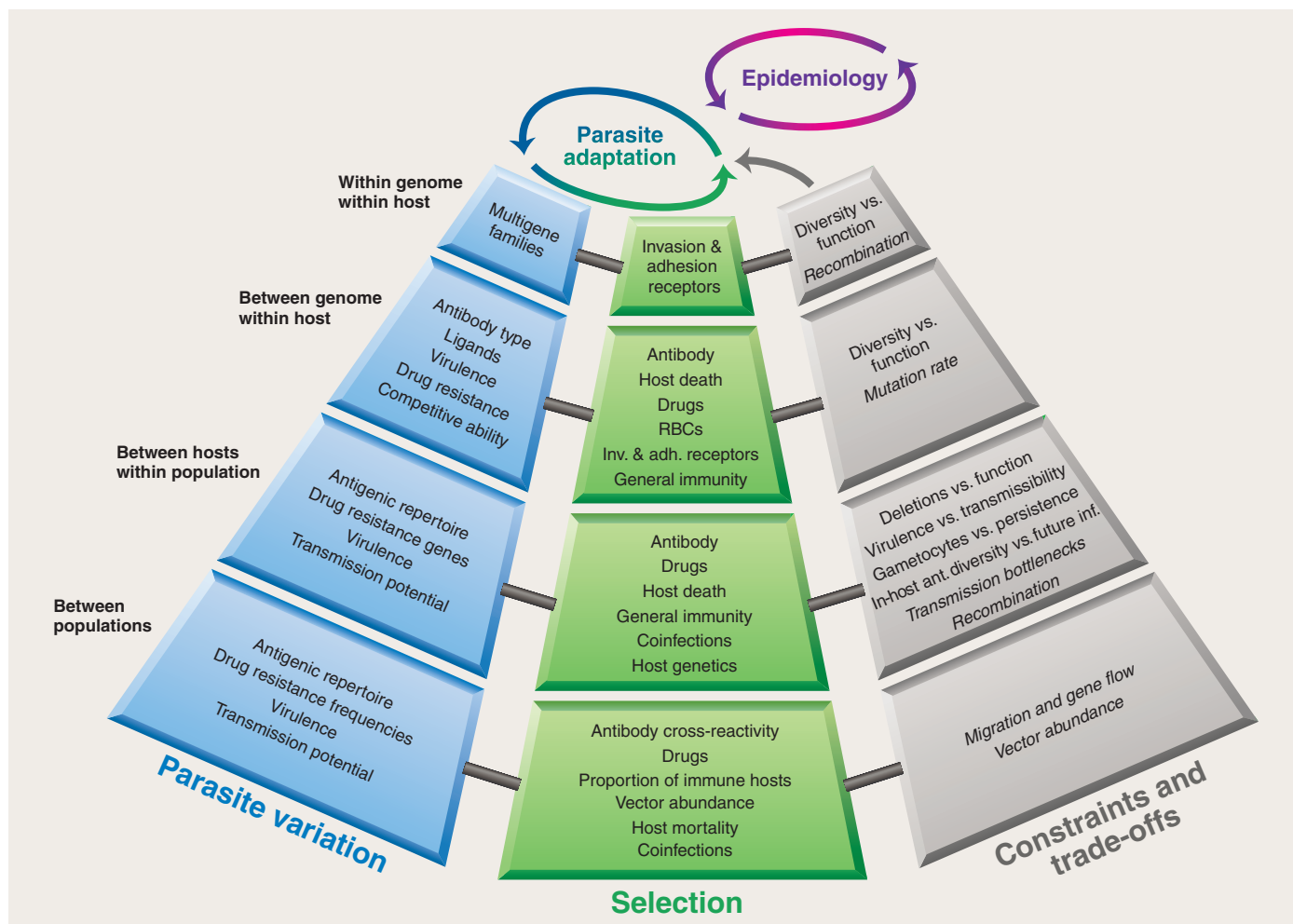


Fig. 1. The selection landscape of *P. falciparum*. Parasite adaptation depends on sources of variation (blue) and selection (green), which, when combined, produce higher parasite fitness. Constraints and trade-offs (gray shapes) impose limits on the adaptive process (gray arrow). All three of these components differ across the multilevel biological, geographical, and temporal scales over which adaptation takes place (indicated on left, one shape per level). There is constant feedback between parasite adaptation and its within-host and epidemiological

environment (red-purple arrows). Words inside the blue shapes describe the types of genes or phenotypes that are variable and on which selection can act. Words inside green shapes describe either the specific selective agent or the general cause of selection on these variants. Constraints are distinguished from trade-offs by italics. Adhesion and invasion receptors are molecules on host cells used by the parasite to survive and replicate and which vary considerably within a host. RBCs, red blood cells; ant., antigenic; inf., infection.

The second example is drug resistance. Typically, point mutations within a single gene are maintained at very low population frequencies before any drug is introduced; these are then selected by the drug, which halts replication of the parasites lacking such mutations. The fitness advantage to being resistant is uninterrupted parasite replication and, hence, transmission. However, resistance brings with it a higher risk that the host dies as a result of the parasite's activities, which is then very costly to parasite fitness. In this case, selection is directional, long-term, and between-host. There may be weak, indirect feedback from the evolution of resistance into the selective environment, leading to a reduction in drug usage, due either to perceived drug failure or to higher immunity caused by higher transmission intensities. A possible constraint is the cost of the mutation to the parasite's fitness in the absence of the drug (10).

In reality, these examples are more complex. Although switching of particular VSAs can be seen primarily as a short-term adaptation, when viewed at the population level and over a longer time scale, selection is operating on the genome's whole repertoire of antigenic and cytoadherence types to maximize parasite fitness. Moreover, selection optimizes the level of population-wide diversity and within-host switching rates so that the parasite can establish infections in previously exposed hosts (9, 11). Thus, there are multiple layers and scales of selection operating on VSAs.

Likewise, during the evolution of drug resistance, selection is not all between-host and long-term. There is a critical stage at the very beginning of the adaptation process during which within-host, short-term processes entirely dictate whether a resistance genotype will emerge (i.e., within-host parasite population size, mutation rate, and

size of population bottlenecks during mosquito transmission) but that are unimportant to its rate of spread once it is safely established in a population (12).

Hence, for the malaria parasite, the process of adaptation to a single environmental challenge takes place across a range of temporal, biological, and geographic scales, each with different selection pressures and outcomes. In the following, we describe in more detail the factors involved and the possible implications for malaria control.

Parasite Variation: Past Adaptation and Future Substrate

The consequences of past adaptation have been "written into" the parasite's genome through evolutionary time, leaving clues that are useful for understanding the selective forces and adaptive mechanisms of the parasite. First, high DNA

sequence variability coupled with a bias for non-synonymous mutations is an indicator of diversifying selection and thereby flags the targets under immune or other variable selection pressures (13). For instance, repetitive units may signal immune evasion mechanisms (14, 15), as might genetic mosaicism arising from nonhomologous recombination among different members of multi-gene families (16), although the maintenance of genetic variation through neutral processes in most genomes also needs to be borne in mind. In contrast, low sequence diversity in a parasite gene and high linkage disequilibrium around the locus indicate directional (“purifying”) strong selection on a gene inside that region. This phenomenon has retrospectively confirmed drug resistance loci found using other genetic methods and has identified several new loci potentially involved in resistance (17, 18).

Second, deletions and amplifications of short DNA segments containing a single gene or a few genes, that is, copy number variants (CNVs), are highly prevalent in natural populations of

P. falciparum (19, 20, 21), indicating that such structural variation may be a substrate for adaptation. So far, amplification CNVs have been linked to resistance to three different classes of drugs (19, 22–24). In the laboratory, certain deletion CNVs containing genes involved in gametocyte production and cytoadherence consistently arise during adaptation to in vitro culture (25–27). The fact that the fragility of certain chromosomal regions is maintained in nature, despite possible in vivo fitness costs, may suggest that these deletion CNVs are adaptive under some circumstances (25). They may provide a growth advantage during the acute phase of the infection when there is strong in-host competition for host cells. In this case, although the deletion-carrying parasite lineage itself may not be transmitted, its kin would be, thereby gaining an advantage over parasite lineages with nonfragile chromosomes. Studies of genome-wide transcription in *P. falciparum* have found evidence that CNVs are associated not only with genes inside the CNV boundaries but also with regulation of genes adjacent to

them and elsewhere in the genome (21, 28). Perhaps the CNVs’ structural properties per se influence gene expression through short and long-range epigenetic mechanisms, as seen in other organisms (29). Structural variation in noncoding regions of the *P. falciparum* genome such as the telomeres is also a potential source of regulatory variation (25).

A third set of clues for adaptive genes lies in genomic architecture. Most of the “exportome” genes, that is, those responsible for transporting parasite proteins out to the red-cell surface, where they bind to host tissues and mediate immune evasion, are located in the subtelomeres (30). There, they are at least partially regulated through heterochromatin-related mechanisms (31, 32). Epigenetic regulation of physical clusters of functionally related genes would facilitate parasite adaptation within an infection where the host environment is continuously changing. Likewise, location of the highly variable multigene families in subtelomeres, which group together in bouquets during mitosis, would provide opportunity for

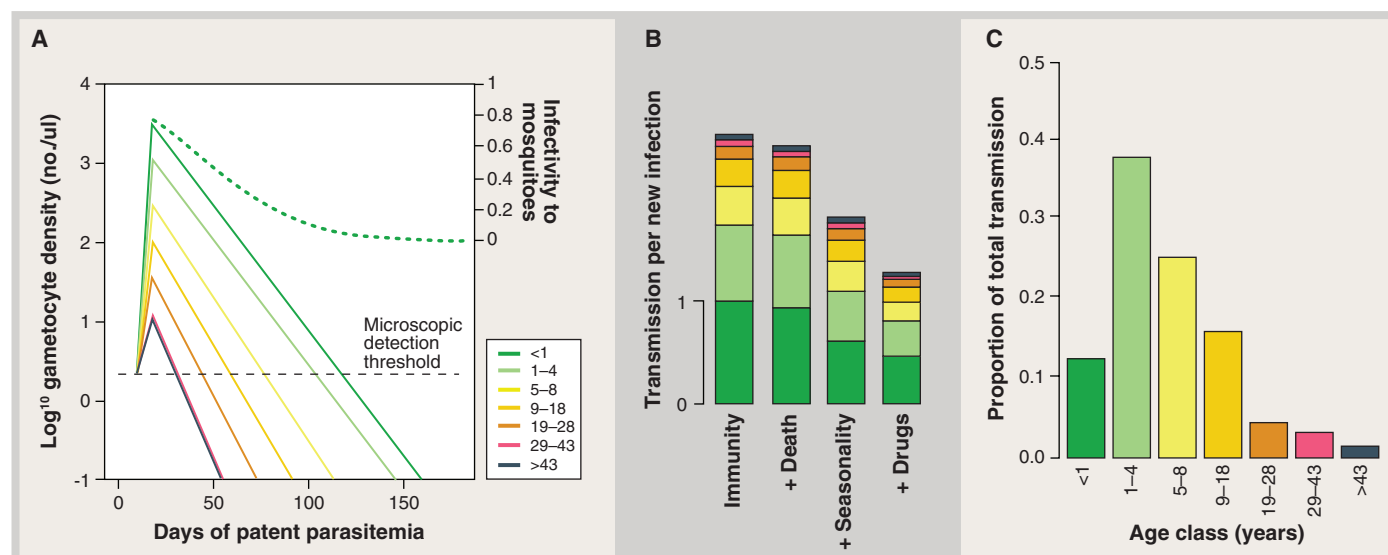


Fig. 2. Quantifying fitness costs of immunity, host death, drug use, and seasonality in *P. falciparum* parasites in the field. **(A)** Total transmission of the malaria parasite from each new human infection can be calculated from the area under the within-host gametocyte density by time curves (solid colored lines) weighted by the probability of mosquito infection given the gametocyte density (dotted line, upper right corner). Different colored curves (legend on right) are given for hosts of different age classes. These were derived from data (52) from cross-sectional surveys in a malaria-endemic area before drugs, bednets, and vector control were introduced (41). Mosquito infectivity (dotted line) is shown for <1-year-olds only. **(B)** Based on these curves, the expected impacts of host immunity, host death, seasonality in vectors, and antimalarial drug usage on total transmission output per new parasite infection per host were calculated for each age class. Expressing all values relative to the total transmission output for <1-year-olds (set to equal 1), the fitness cost to the parasite of age-acquired immunity (first bar, reading from left) ranged from 26% in 1- to 4-year-olds to 95% in >29-year-olds (i.e., parasites in these age groups were expected to have 74% and 5% of the transmission output of parasites infecting <1-year-olds). The cost of host death (second bar) was predicted to be 3% of total transmission in <1-year-olds and 2.5% in 1- to

4-year-olds, assuming that all-cause mortality in <5-year-olds was 120 per 1000 person-years and that mortality was distributed among age classes according to age-mortality patterns typical of endemic areas (53). The impact of seasonality in vector abundance (third bar), assuming that transmission ceased on average 45 days after hosts became patent, was predicted to be 40% of transmission in <1-year-olds and relatively less at older ages. Drug treatment (fourth bar), when assumed to cause complete loss of transmission and to follow age usage patterns from a rural malaria-endemic area in Mozambique (54), was predicted to cut transmission by 54% in <1-year-olds and 44% in >29-year-olds. **(C)** Because the age-specific impact of these costs to the total transmission affects the selection pressures on the parasite coming from each host age class, the relative contributions of each age class to total transmission are shown. The latter calculations were based on the area under the transmission curve for total infection length (i.e., including superinfections) multiplied by both the infection prevalence and the proportion of the population in each age class. The lower contribution to transmission in <1-year-olds than in 1- to 4-year-olds is due to their lower prevalence and shorter total infection lengths due to fewer superinfections (39) rather than to their proportionately smaller number in the population.

recombination among different family members and hence promote antigenic diversity (33).

DNA sequence diversity alone belies the extent of diversity at the molecular level because it does not reveal regulatory variation, that is, in transcription and translation. A large amount of regulatory variation between different parasite isolates (21, 28, 34–36) occurs in genes showing no genetic diversity (21, 35). Some of this variation may be caused by temporary conditions in the host, such as drug treatment (37) or physiological status [e.g., oxidative stress or pregnancy (35, 38)] that will influence transcriptional pathways in the parasite positively or negatively. Defining the relative contributions of genetically fixed between-strain differences, epigenetics, and temporary environmentally induced variability in transcription and translation is of major importance to the question of how much phenotypic variation in malaria parasites is heritable and passed intact from host to host via the mosquito, and therefore contributes to evolutionary change.

Ultimately, selection acts on whole-organism phenotypes (traits), such as asexual replication rate, transmissibility, persistence, and virulence. Combined, these traits determine the parasite's fitness, that is, the total number of successful transmissions to a new host that the parasite achieves within the duration of the infection of its current host, roughly calculated as the rate of transmission (transmissibility) multiplied by the duration of infection (persistence). Virulence, defined here as the propensity to cause host death, shortens the duration of infection, and so acts negatively on parasite fitness. Studies in a laboratory model of malaria, *P. chabaudi* in mice, have shown that the intrinsic asexual replication rate of a parasite strain is the primary determinant of transmissibility, persistence, and virulence, and therefore underpins parasite fitness (39, 40). Similarly, in natural infections of *P. falciparum*, asexual parasite density is related to transmissibility and persistence (39, 41), although the contributions of parasite strain and host immunity cannot be separated in this case.

Is there parasite genetic variation in these traits on which selection could act? *P. falciparum* strain variation has been observed in asexual replication rate and gametocyte production both in vitro and in vivo in the growth of lines that have been passaged through humans (39). Thus, the potential, at least, for there being heritable variation in fitness component traits has been demonstrated, even if its quantity in natural populations has not.

What Are the Major Selective Pressures on the Parasite?

Among the many sources of selection acting on the malaria parasite, five dominate: immunity, vector availability, host death, drugs, and coinfection. If the host dies (which occurs in about 1% of cases), it usually does so during the acute phase of the illness before the transmissible forms

Box 1. Constraints and trade-offs.

Most traits do not evolve in an unlimited way because they are constrained by factors that either impede genetic change (e.g., lack of variability or gene flow) or because of competing functions or selective forces (trade-offs).

Constraints.

- Population bottlenecks during transmission of malaria parasites are severe, thus wasting most mutational variation generated within each infection of each host species. Nevertheless, genome-wide surveys indicate that parasite genetic diversity is very high (17, 21) and unlikely to be limiting to rates of adaptation.

Trade-offs.

- If selection for transmission potential also selects for virulence (probability of host death), this will generate a trade-off because parasites cannot achieve maximum fitness when virulence and transmission potential are both high.

- For every gametocyte produced, an asexual lineage has to sacrifice its future asexual reproduction and hence future transmission potential. The costs of high *Plasmodium* infection to the mosquito also may trade against gametocyte production (55, 56).

- Shifting patterns of antigenic diversity in VSAs under immune selection suggest that there is some trade-off between their function both as antigens and as cytoadherence ligands, i.e., virulence factors (6, 7).

- If infection with *Plasmodium* reduces fitness of the mosquito, and does so disproportionately more in the more virulent parasite strains, there will be a trade-off between virulence and transmission, although no strong relationship has yet been observed between virulence in the vertebrate and mosquito hosts (57).

- The frequency of chloroquine-resistant genotypes in a population sometimes decreases after chloroquine is withdrawn from use (14), which suggests that there is a metabolic cost to chloroquine resistance in the absence of the drug. Whether or not the maintenance of resistance in some regions is due to compensatory changes in asexual replication ability, transmissibility, or infection length is not yet established.

(gametocytes) appear in the bloodstream, thus causing the parasite to lose all of its transmission potential (Fig. 2A). Some drugs, if used effectively, have the same effect. Host immunity reduces transmission potential by curtailing asexual replication, thereby reducing both gametocyte production (transmissibility) and the duration of infection (persistence). Ineffectively used drugs can do the same. Immunity and drugs, when effective, can prevent host death and rescue the parasite. Thus, there are both fitness benefits (avoiding host death) and costs (reduced transmission potential) of immunity and drug use. By contrast, a loss of vector availability is always costly to the parasite: When the population density of the mosquito vector is reduced by seasonal factors, the fitness cost is from curtailed infections, whereas in non-seasonal transmission areas, the fitness cost is from reduced transmissibility.

By quantifying the realized costs of host death, immunity, drugs, and seasonality in *P. falciparum* infections under the typical conditions prevailing in a high-transmission, malaria-endemic area, each of these selection pressures alone is shown to have a substantial effect, with immunity being the most detrimental and host death the least detrimental at the population-wide level (Fig. 2B). Although the quantitative outcome of this analysis may vary with transmission setting, it reveals several important principles about parasite evolution. First, the costs vary across age and immunity classes in the host population; for example, the

cost of host death is greatest in children because they are more likely to die and also transmit most, and is negligible in adults. Second, age differentially affects the various types of costs: Parasites from adult hosts lose more fitness due to immunity than from host death. The consequence of these two principles is that, because the benefit-cost ratios of each selective force are age and immunity dependent, the level of the trait (e.g., virulence) that is optimal (i.e., balances the benefits and costs) in one class of hosts will be suboptimal in another. As the parasite has to evolve to suit the average host because it cannot choose which host type it infects, what matters for the optimum level is the relative population sizes of these immunity classes in the host population. For example, children may suffer more death than is optimal for the parasite because immune adults in the population have driven the parasite's virulence optimum higher. Third, selection on the parasite in children (in this example) would be more effective in bringing about evolutionary change because children contribute most to the total transmission pool (Fig. 2C). Fourth, these selection forces cannot be considered in isolation because they often interact. For example, as immunity increases, host mortality is likely to decrease, and drugs may also be used less often in semi-immune (where infections are typically asymptomatic) than nonimmune people. Finally, by affecting the transmission output of the parasite, all these forces alter the transmission intensity in the environment and

hence the age and immunity structure of the host population, which feeds back into the parasite's evolutionary optima. Thus, to get a complete picture of the parasite's adaptation in response to interventions, an understanding is required of the epidemiological dynamics, as well as the fitness costs and benefits involved in individual parasite infections.

Other costs to the parasite's fitness can arise from coinfection of hosts with other pathogen species and alternative genotypes of the same species. In malaria-endemic areas, infection with multiple parasite clones is common (42). Parasites of the same species can compete in two ways, either by direct competition for host resources, such as red blood cells, or indirectly for "immunity-free space" (42, 43). A third way in which coinfecting parasites impact on each other's fitness is through virulence; that is, if one strain is too virulent and kills its host, all coinfecting strains suffer loss of absolute fitness (a "tragedy of the commons" scenario). The collective evidence from controlled coinfection experiments in the mouse malaria model, *P. chabaudi*, suggests that all these potential effects of coinfection are important (42–45). How does competition affect the direction of selection of parasite fitness traits? A tentative generalization owing to the small number of studies and complexity of within-host competitive interactions (42) is that parasites with the fastest intrinsic growth rates are the ones that have the highest competitive ability in terms of asexual growth, are the most virulent, and have greater transmission output (43–45).

Another emerging generalization is that immunity exacerbates the strength of competition (43, 46). If both these generalizations apply to *P. falciparum* in its natural setting, it is predicted that there will be a fitness advantage to parasites with high intrinsic growth rates because they compete better in terms of asexual growth and transmission potential. They will be at an advantage in high transmission areas where mixed infections are more common and where immunity levels are higher. However, there is much left to learn about within-species competition in *P. falciparum* infections and other parasitic diseases and its consequences to public health. Similarly, when another pathogen species coinfects the host, the fitness of the malaria parasite *P. falciparum* will also be affected. Because each combination of pathogen species will elicit a different set of interactions, these will need to be studied on a case-by-case basis to predict the likely outcome of control of the other pathogen.

Which Direction Will the Selection Response Take?

Given that the fitness costs to the parasite of immunity, drugs, vector availability, and host death are substantial, and that the benefit-cost ratios of fitness traits differ across host age and immunity

classes, it follows that changes in environmental conditions or control interventions that alter the transmission benefits and costs will generate evolutionary change in the parasite population. This idea has been most explored for virulence. For example, it has been argued that widespread use of an asexual-stage vaccine producing good but incomplete immunity might select for higher virulence because the most virulent parasites can more often "get away with it" under the cover of predominantly semi-immune hosts (47). Conversely, it is expected that as host populations become less immune due to vector control, infection-blocking liver-stage vaccines, and bed-nets, there will be evolution to lower virulence because the highly virulent parasites will be removed by host death. A similar argument has been made for drug use in the absence of resistance (48) and could also be made for host resistance genes. This is because, like immunity, drugs and host resistance stop host death and allow virulent parasites to thrive, but they also cut transmission, which the parasite must then recover by evolving higher transmissibility and persistence. In these arguments, it is assumed that parasites with higher virulence have, on average, higher transmission potential than their less virulent counterparts, which leads to a higher frequency of virulent parasites once the cost of host death is removed.

What of other traits, such as persistence or transmissibility? Independently of their relationship with virulence, they are expected to evolve to ever greater levels. The fact that they have not done so in malaria parasites (i.e., gametocyte densities are maintained at relatively low densities in the blood, and infections do not last for a host's lifetime) suggests that there are some major constraints operating on these traits (Box 1).

One trait that clearly could evolve independently of virulence is antigenic diversity. The question of how it should evolve optimally for the parasite can be answered by following the fitness benefit-cost principles discussed in the previous section. Does a novel antigen increase or decrease the probability of host death? To what extent is the cost of host death offset by the gain or loss in transmission potential? Does the antigenic novelty come at a price in terms of function and fitness? How do these costs and

benefits alter across hosts with different levels of immunity? How do they alter within an infection as immunity changes? So far, the evidence suggests that, in the case of *var*-encoded surface antigens, antigenic diversity does not appear to evolve independently of virulence owing to competing selective forces on antigenic novelty and cytoadherence function (Box 1) (6, 7).

This fitness benefit-cost approach does not work for all situations because traits and genes do not evolve in isolation, nor are populations always at epidemiological and evolutionary equilibria. Nevertheless, taking this approach is a useful exercise for deciding which selection pressures are most likely to drive evolutionary change and in what direction.

How Long Will It Take and What Should We Do?

Success in malaria control is proceeding apace, but not even the most optimistic expect malaria to be eradicated in a few decades. The experience with drug resistance indicates that rates of parasite evolution can be fast: All mass-deployed drugs have developed resistance within a few years of use and failed widely within one to three decades of use (49). This may be so extreme because of the very high selective pressures that drugs impose (Fig. 2B), but calculations of rates of evolution of virulence under a partially effective and partially deployed blood-stage vaccine similarly suggest it might take 20 to 30 years to evolve from a prevaccination to a postvaccination level of virulence (47). Thus, parasite evolution will erode the efficacy of the control measure, and it is not a matter of elimination or evolution but of which wins the race. This will depend on the relative impact of the control measure versus that of parasite adaptation on the epidemiology of the disease.

Control interventions need to take into account the direct benefits to the individual person (e.g., a drug will stop a host dying), the indirect benefits to the population (e.g., widespread drug use will reduce transmission intensity), and the evolutionary consequences to the parasite population that might negate these benefits (e.g., drug pressure will select for resistance and ultimately render the drugs useless). In presenting a general framework for malaria

Box 2. Principal issues for future research.

- How fast and how widely are natural populations of malaria parasites adapting in parasite fitness traits such as virulence, transmissibility, and persistence?
- How much within-population standing variation is there in parasite fitness traits?
- How much of the phenotypic variation in these traits is due to DNA sequence diversity, structural variation in the genome, regulatory variation, epigenetics, and temporary environmental or host influences, and how heritable is it?
- Which parasite molecular mechanisms are associated with parasite fitness?
- Are geographically separated parasite populations differentially adapted to their environments?
- What are the genetic variants underlying adaptation to different environmental conditions?

parasite evolution that centers around fitness costs and benefits, and how these vary across and are shaped by different environmental conditions (e.g., population levels of immunity), we hope to facilitate thinking on circumstances under which control programs should hit hard or tread gently, prevent or cure, use mass or targeted interventions, and prioritize public health over private benefit.

A challenging example of this is the policy decision that needs to be made about what stage in an elimination program drugs should be used to cure asymptomatic infections versus only symptomatic ones. The individual person will be largely unaffected by this decision, whereas the population will benefit as a whole if treatment is widespread. But if there is even a very low frequency of resistance in the population, widespread drug use in all parasitemic individuals may hasten the spread of resistance by increasing the selection pressure on the parasite.

These conflicts between epidemiological control and evolutionary erosion of control efficacy, although largely inevitable, can be mitigated, and we need to identify the strategies that work best for both the short-term and long-term. An example of this is the choice to use control agents that kill *Plasmodium*-carrying adult mosquitoes when they are old rather than when they are young: This is because while the effect on parasite transmission (and hence public health) is the same regardless of the age the mosquitoes die, the efficacy of selection for resistance to these control agents is much lower in older mosquitoes because their reproductive potential has been almost exhausted (50). Preventing gene flow between parasite populations carrying undesirable properties [e.g., high virulence and/or drug resistance (51)], especially in areas where elimination is within reach, is another example of how the consequences of interventions can be controlled. Often, however, the ability to make these sorts of recommendations depends on knowledge that we do not yet have.

Because parasite fitness is so difficult to measure, future research to address outstanding issues in malaria parasite adaptation (Box 2) will probably require indirect approaches, such as measuring differences between parasite populations that have evolved under different environmental conditions or measuring properties of parasites within populations after they have been filtered by selection (e.g., parasites in immune versus nonimmune hosts, hosts with severe versus mild disease, or hosts treated with drugs versus untreated). Genomics-related technologies that can simultaneously assess all the genetic and molecular variation, such as new-generation sequencing, microarray-based measures of whole transcriptomes, and high-throughput proteomics,

will allow a more direct and faster route from phenotype to genotype than ever before. By creating a better understanding of gene regulatory and functional networks in the parasite, it is hoped that a more holistic approach to understanding how *P. falciparum* maintains its extraordinary transmission capacity may help identify new and durable targets for parasite control.

References and Notes

- R. W. Snow, C. A. Guerra, A. M. Noor, H. Y. Myint, S. I. Hay, *Nature* **434**, 214 (2005).
- K. H. Rieckmann, *Ann. Trop. Med. Parasitol.* **100**, 647 (2006).
- M. J. Gardner *et al.*, *Nature* **419**, 498 (2002).
- A. G. Maier, B. M. Cooke, A. F. Cowman, L. Tilley, *Nat. Rev. Microbiol.* **7**, 341 (2009).
- S. A. Kyes, P. Horrocks, C. I. Newbold, *Annu. Rev. Microbiol.* **55**, 673 (2001).
- P. C. Bull *et al.*, *J. Infect. Dis.* **182**, 252 (2000).
- G. M. Warimwe *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 21801 (2009).
- S. Gupta, K. Trenholme, R. M. Anderson, K. P. Day, *Science* **263**, 961 (1994).
- M. Recker, N. Arinaminpathy, C. O. Buckee, *Malar. J.* **7**, 18 (2008).
- J. G. Kublin *et al.*, *J. Infect. Dis.* **187**, 1870 (2003).
- C. O. Buckee, P. C. Bull, S. Gupta, *Proc. Biol. Sci.* **276**, 477 (2009).
- M. J. Mackinnon, *Acta Trop.* **94**, 207 (2005).
- J. Mu *et al.*, *Nat. Genet.* **39**, 126 (2007).
- R. F. Anders, D. J. McColl, R. L. Coppel, *Acta Trop.* **53**, 239 (1993).
- A. L. Hughes, *J. Mol. Evol.* **59**, 528 (2004).
- M. Frank *et al.*, *Int. J. Parasitol.* **38**, 1099 (2008).
- S. K. Volkman *et al.*, *Nat. Genet.* **39**, 113 (2007).
- J. Mu *et al.*, *Nat. Genet.* **42**, 268 (2010).
- C. Kidgell *et al.*, *PLoS Pathog.* **2**, e57 (2006).
- U. Ribacke *et al.*, *Mol. Biochem. Parasitol.* **155**, 33 (2007).
- M. J. Mackinnon *et al.*, *PLoS Pathog.* **5**, e1000644 (2009).
- N. V. Dharia *et al.*, *Genome Biol.* **10**, R21 (2009).
- S. Nair *et al.*, *PLoS Genet.* **4**, e1000243 (2008).
- T. E. Welles *et al.*, *Nature* **345**, 253 (1990).
- L. G. Pologe, J. V. Ravetch, *Cell* **55**, 869 (1988).
- B. A. Biggs, D. J. Kemp, G. V. Brown, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2428 (1989).
- K. P. Day *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 8292 (1993).
- J. M. Gonzales *et al.*, *PLoS Biol.* **6**, e238 (2008).
- S. Henikoff, *Genetics* **138**, 1 (1994).
- M. Marti, R. T. Good, M. Rug, E. Kneuper, A. F. Cowman, *Science* **306**, 1930 (2004).
- J. J. Lopez-Rubio, L. Mancio-Silva, A. Scherf, *Cell Host Microbe* **5**, 179 (2009).
- C. Flueck *et al.*, *PLoS Pathog.* **5**, e1000569 (2009).
- L. H. Freitas-Junior *et al.*, *Nature* **407**, 1018 (2000).
- M. Llinás, Z. Bozdech, E. D. Wong, A. T. Adai, J. L. DeRisi, *Nucleic Acids Res.* **34**, 1166 (2006).
- J. P. Daily *et al.*, *Nature* **450**, 1091 (2007).
- H. Jiang *et al.*, *PLoS ONE* **3**, e2484 (2008).
- G. Hu *et al.*, *Nat. Biotechnol.* **28**, 91 (2010).
- S. E. Francis *et al.*, *Infect. Immun.* **75**, 4838 (2007).
- M. J. Mackinnon, A. F. Read, *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **359**, 965 (2004).
- R. Timms, N. Colegrave, B. H. K. Chan, A. F. Read, *Parasitology* **123**, 1 (2001).
- L. Molineaux, G. Gramiccia, *The Garki Project: Research on the Epidemiology and Control of Malaria in the Sudan Savanna of West Africa* (World Health Organization, Geneva, 1980).
- A. F. Read, L. H. Taylor, *Science* **292**, 1099 (2001).
- L. Raberg *et al.*, *Am. Nat.* **168**, 41 (2006).
- J. C. de Roode *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 7624 (2005).
- A. R. Wargo, S. Huijben, J. C. de Roode, J. Shepherd, A. F. Read, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 19914 (2007).
- M. J. Mackinnon, A. F. Read, *PLoS Biol.* **2**, e230 (2004).
- S. Gandon, M. J. Mackinnon, S. Nee, A. F. Read, *Nature* **414**, 751 (2001).
- T. C. Porco, J. O. Lloyd-Smith, K. L. Gross, A. P. Galvani, *J. Theor. Biol.* **233**, 91 (2005).
- C. Wongsrichanalai, A. L. Pickard, W. H. Wernsdorfer, S. R. Meshnick, *Lancet Infect. Dis.* **2**, 209 (2002).
- A. F. Read, P. A. Lynch, M. B. Thomas, L. D. Hurst, *PLoS Biol.* **7**, e1000058 (2009).
- A. M. Dondorp *et al.*, *Nat. Rev. Microbiol.* **8**, 272 (2010).
- Average gametocyte density curves for each age class were predicted from average asexual densities and infection lengths observed in cross-sectional surveys in the Garki Project (41), assuming a log-linear rate of decline from peak parasite density on day 8 until the threshold of detection by microscopy was reached, and a 10-day delay between peak asexual density and gametocyte density. This curve shape was assumed because it gave good fit to daily data on asexual densities during *P. falciparum* infections used to treat neurosyphilis (58–60). Gametocyte densities were calculated from asexual densities based on the linear relationship between the age-class averages of these traits in the Garki Project (39, 41). Because the focus here is on parasite fitness, the curves shown in Fig. 2A are for a single parasite infection, that is, in the absence of superinfection that renders actual infection lengths much longer than these, especially in the 1- to 4- and 5- to 8-year-olds where parasite infections accumulate faster than they are cleared (39, 41). The relationship between mosquito infectivity and average gametocyte density in Fig. 2A was calculated from data on experimental feeding of mosquitoes to *Plasmodium*-infected patients (60). Because host death, when it occurs, does so soon after peak asexual parasite density before gametocytes infectious to mosquitoes have appeared, no differential weighting of transmissibility to account for changing host survival probability through time was applied.
- I. Carneiro *et al.*, *PLoS ONE* **5**, e8988 (2010).
- R. E. Mota, A. M. Lara, E. D. Kunkunzu, D. G. Lalloo, *Am. J. Trop. Med. Hyg.* **81**, 935 (2009).
- L. H. Taylor, A. F. Read, *Parasitol. Today* **13**, 135 (1997).
- N. Mideo, T. Day, *Proc. Biol. Sci.* **275**, 1217 (2008).
- H. M. Ferguson, M. J. Mackinnon, B. H. K. Chan, A. F. Read, *Evolution* **57**, 2792 (2003).
- W. E. Collins, G. M. Jeffery, *Am. J. Trop. Med. Hyg.* **61** (suppl.), 4 (1999).
- W. E. Collins, G. M. Jeffery, *Am. J. Trop. Med. Hyg.* **61** (suppl.), 44 (1999).
- W. E. Collins, G. M. Jeffery, *Am. J. Trop. Med. Hyg.* **68**, 366 (2003).
- We are grateful to P. Bull, V. Thathy, S. Gandon, and an anonymous reviewer for useful comments. M.J.M. is supported by Wellcome Trust Project Grant 088634, and both authors are supported by Wellcome Trust Core Programme Grant 077092. M.J.M. is a member of the GlaxoSmithKline Biologicals advisory board for modeling malaria vaccines. This paper is published with the permission of the director of the Kenya Medical Research Institute.

10.1126/science.1185410

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

The following resources related to this article are available online at www.sciencemag.org (this information is current as of May 13, 2010):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/328/5980/872>

Supporting Online Material can be found at:

<http://www.sciencemag.org/cgi/content/full/science.1187959/DC1>

This article **cites 29 articles**, 11 of which can be accessed for free:

<http://www.sciencemag.org/cgi/content/full/328/5980/872#otherarticles>

This article has been **cited by 1** articles hosted by HighWire Press; see:

<http://www.sciencemag.org/cgi/content/full/328/5980/872#otherarticles>

This article appears in the following **subject collections**:

Molecular Biology

http://www.sciencemag.org/cgi/collection/molec_biol

Small Silencing RNAs in Plants Are Mobile and Direct Epigenetic Modification in Recipient Cells

Attila Molnar,* Charles W. Melnyk,* Andrew Bassett, Thomas J. Hardcastle, Ruth Dunn, David C. Baulcombe†

A silencing signal in plants with an RNA specificity determinant moves through plasmodesmata and the phloem. To identify the mobile RNA, we grafted *Arabidopsis thaliana* shoots to roots that would be a recipient for the silencing signal. Using mutants that block small RNA (sRNA) biogenesis in either source or recipient tissue, we found that transgene-derived sRNA as well as a substantial proportion of the endogenous sRNA had moved across the graft union, and we provide evidence that 24-nucleotide mobile sRNAs direct epigenetic modifications in the genome of the recipient cells. Mobile sRNA thus represents a mechanism for transmitting the specification of epigenetic modification and could affect genome defense and responses to external stimuli that have persistent effects in plants.

In flowering plants, RNA silencing is a non-cell-autonomous process: It spreads both to neighboring cells and systemically over long distances (1–3). Mobile silencing operates in a nucleotide sequence-specific manner consistent with the involvement of RNA as a component of the signal, but until now the mobile RNA has not been detected. The 24-nucleotide (nt) and 21-nt

species of small RNA (sRNA) have both been implicated in mobile silencing, but it could not be ruled out that a long sRNA precursor is the form that moves (4, 5). Conversely, analyses with grafted plants suggested that long sRNA precursors are mobile (6), although the data did not rule out sRNAs. Here, we describe a similar grafting approach with *Arabidopsis thaliana*

combined with high-throughput sequencing to detect 21- to 24-nt sRNA molecules that had moved across the graft union. Silencing pathway mutants allow us to demonstrate that the mobile sRNA is synthesized in the source rather than the recipient tissue. We show that mobile 24-nt sRNAs from three separate genomic loci can direct epigenetic modifications in the recipient cells.

Transgene-specific sRNAs are mobile. The grafting experiments used transgenic *A. thaliana* plants expressing a green fluorescent protein (GFP) transgene (G) as a reporter of RNA silencing (7) and a GFP-derived hairpin RNA [GF inverted-repeat (GF-IR)] to silence GFP (Fig. 1A). GFP-expressing plants are green under ultraviolet light, so that the movement of the silencing signal can be monitored through the loss of green fluorescence (Fig. 1, B and C, and fig. S1, A to F). In test grafts we ascertained that movement of the GFP silencing signal was more efficient from shoot to root than vice versa, consistent with previous findings of the source-to-sink movement of viruses and assimilates (8, 9) (fig. S1, C and E). Subsequent experiments to analyze

Department of Plant Sciences, University of Cambridge, Cambridge CB2 3EA, UK.

*These authors contributed equally to this work.

†To whom correspondence should be addressed. E-mail: dcb40@cam.ac.uk

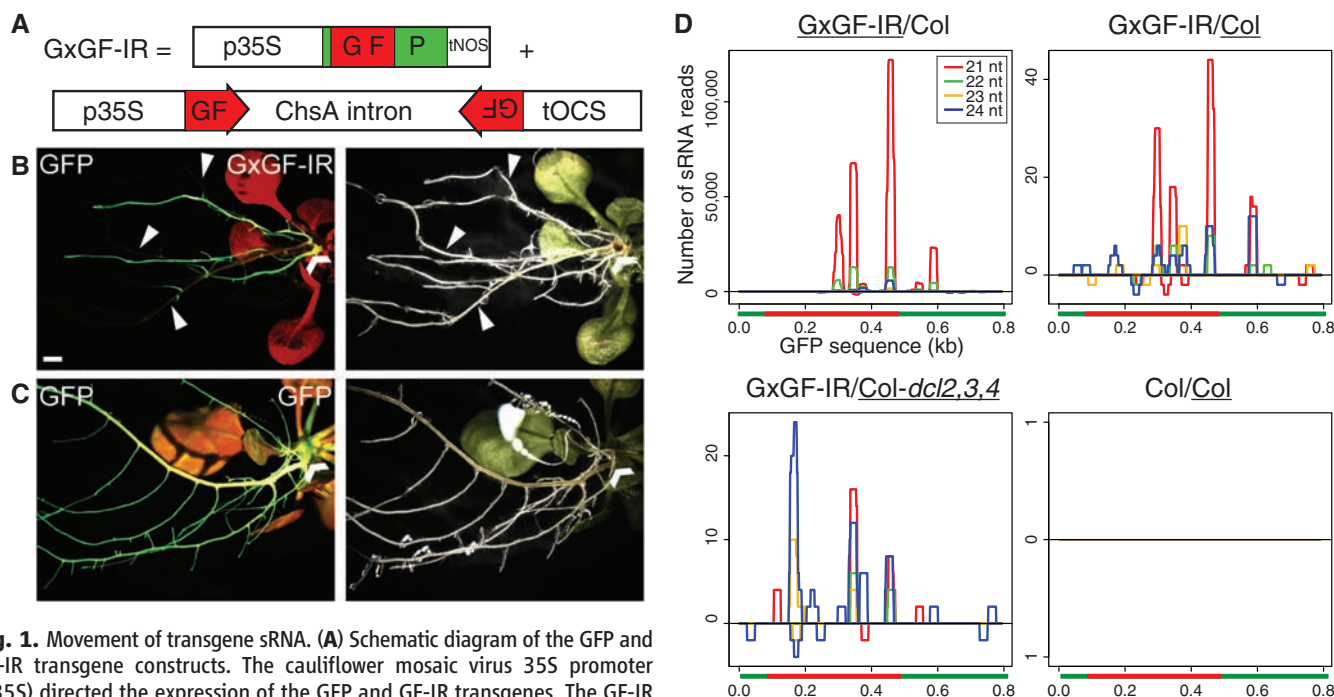


Fig. 1. Movement of transgene sRNA. (A) Schematic diagram of the GFP and GF-IR transgene constructs. The cauliflower mosaic virus 35S promoter (p35S) directed the expression of the GFP and GF-IR transgenes. The GF-IR construct contained a chalcone synthase A (ChsA) intron. The GFP and GF-IR transgenes contained nopaline synthase (tNOS) or octopine synthase (tOCS) terminators, respectively. (B and C) GFP-silenced [GxGF-IR (B)] or GFP-expressing (C) shoots were grafted onto GFP-expressing roots of 7-day-old *Arabidopsis* seedlings, and ultraviolet fluorescence (left) and bright-field images (right) were taken 5 weeks after grafting. Arrowheads indicate emerging silenced roots; chevrons indicate graft junction. Scale bar, 0.5 cm.

(D) sRNA sequence libraries generated from grafted *Arabidopsis* were aligned at the first nucleotide position to the GFP coding sequence; the positive or negative y axis shows the number of reads at each position on either plus or minus strand. The line below the x axis represents positions of the GFP coding sequence (green) and the inverted-repeat (GF) region (red). See fig. S1 for an independent biological replicate, and table S1 for a summary of sRNA library details.

mobile silencing were therefore carried out to assay shoot-to-root movement.

sRNA cDNA libraries were generated from signal source and recipient tissues 5 weeks after grafting. (The sampled tissue in grafted samples is underlined below using shoot/root notation.) The receiving root tissue was from nontransgenic Columbia (Col) genotype plants or from *Col-dcl2,3,4* triple mutant plants that are unable to

produce 22-, 23-, and 24-nt sRNAs from double-stranded RNA precursors (10). In the signal donor GxGF-IR shoots, the lack of GFP expression was associated with the accumulation of GFP-derived sRNAs. All size classes of silencing RNAs (21, 22, and 24 nt) were present, indicating that multiple DCL proteins can act on the precursor of GFP silencing (Fig. 1D). There was a bias toward the accumulation of sense

strand-specific, GFP-derived sRNAs from localized “hotspot” regions of the GFP RNA. This nonrandom distribution of the sRNA could be due to differential processing and/or stability of sRNA from different regions of the GFP transgenes (Fig. 1D), and it is likely a mixture of primary and secondary sRNAs that were produced by an RNA-dependent RNA polymerase on the full-length GFP RNA (11).

The grafted Col roots had GFP-specific sRNAs that were 21 to 24 nt in size, although they were less abundant than in the source tissue by three orders of magnitude (table S1). These RNAs were predominantly from the same hotspots of the sense GFP strand as in the source tissue (Fig. 1D and fig. S1G). A similar population of sRNAs was present in GxGF-IR/*Col-dcl2,3,4* roots, although the proportion of 24-nt RNA relative to the 21-nt sRNAs was higher than in the source tissue or in the Col receiving tissue (Fig. 1D and fig. S1G). These 22- to 24-nt sRNAs could not have been produced in the recipient tissue because of the absence of both DCL2 and DCL3, and hence they must have translocated from the source tissue. These mobile RNAs corresponded to the “GF” and “P” region in both Col and *Col-dcl2,3,4* roots, indicating that both primary and secondary sRNAs are mobile (Fig. 1D and fig. S1G).

Endogenous mobile small RNAs. To find out whether endogenous sRNAs are mobile, we analyzed the size profile of the sRNA libraries from various wild-type and mutant genotypes. The wild-type genotype used in this analysis was the GxGF-IR line in the C24 genotype background (7). The RNA silencing mutants were *Col-dcl2,3,4* in the Col background and a mutant (*sde4-2*) in C24 that lacks the functional RNA polymerase IV (Pol IV) required for 24-nt sRNA biogenesis (12). Col is the most widely used *Arabidopsis* genotype and its genome has been sequenced (13); C24 and Col are more genetically diverse than most other pairwise comparisons of *Arabidopsis* genotypes (14). Because transgene-derived sRNAs accumulate to similar levels in the sRNA libraries compared, the transgenes are not relevant to the analysis of the endogenous sRNAs, and these genotypes are referred to below as Col, C24, *Col-dcl2,3,4*, and *C24-sde4*.

The grafted Col roots accumulated more 23- and 24-nt sRNAs than 21-nt sRNAs, according to an analysis of either total or nonredundant sequence reads (Fig. 2A). The 21-nt species include microRNAs (miRNAs) and small interfering RNAs (siRNAs), whereas the 23- and 24-nt class includes the siRNAs associated with chromatin silencing (15, 16). In the *Col-dcl2,3,4/Col-dcl2,3,4* roots, the 23- and 24-nt sRNAs were reduced to the background level because of the lack of DCL3. However, the 23- and 24-nt sRNAs in the *C24/Col-dcl2,3,4* roots were near wild-type levels, indicating that the endogenous sRNAs are mobile and that they account for a substantial proportion of the endogenous sRNA. These mobile sRNAs are likely

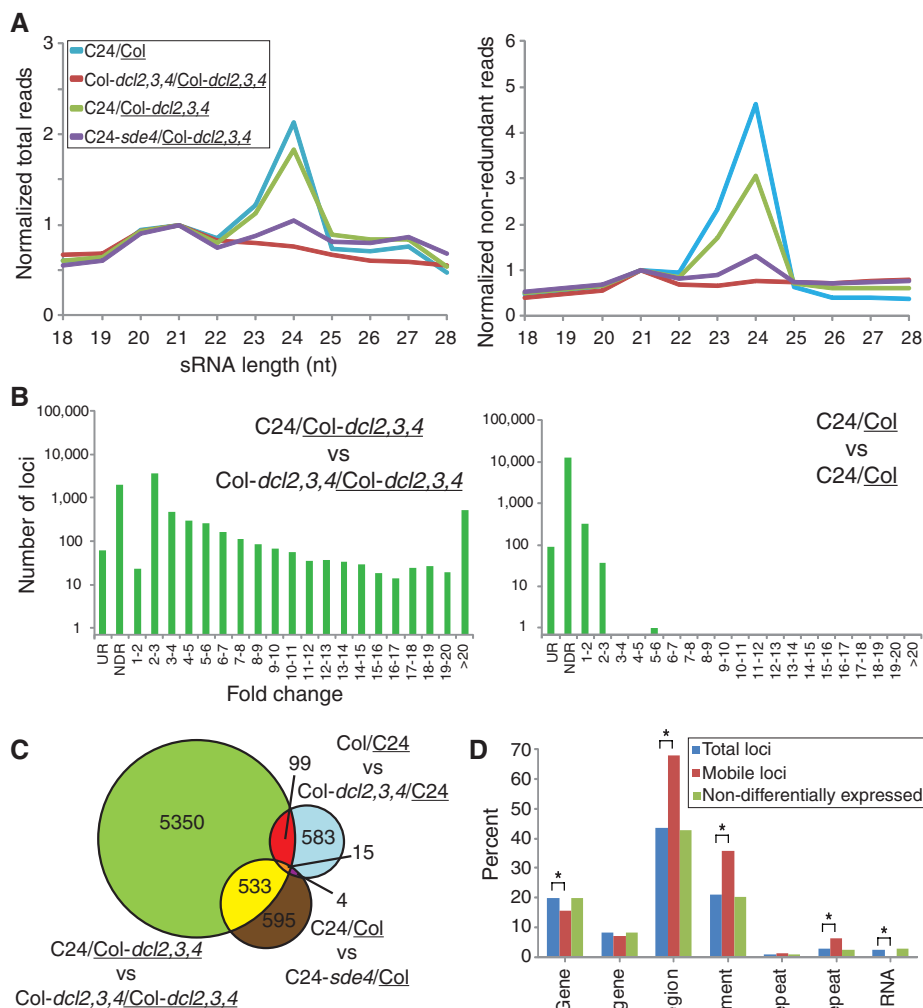


Fig. 2. Many endogenous loci produce mobile sRNAs. (A) Size profiles of total or nonredundant sRNAs from grafted plants, normalized to the number of 21-nt reads. Grafting had a significant effect (log-linear model fitting, $P < 2.2 \times 10^{-16}$) on numbers of sRNAs observed in the size classes for both total and nonredundant reads. (B to D) Loci based on genome-matching total sRNA counts and defined as having a >90% likelihood of expression above background levels, with at least 100 sRNA sequences in one or more of the sRNA libraries used for the pairwise comparison. Mobile sRNA loci are defined as having a >90% likelihood of differential representation. (B) Observed relative changes in mobile sRNA loci overrepresented in *C24/Col-dcl2,3,4* versus *Col-dcl2,3,4/Col-dcl2,3,4*. Replicate *C24/Col* comparisons are shown as a control. Underrepresented (UR) and non-differentially represented (NDR) sRNA loci are shown for comparison. (C) Venn diagram illustrating the extent to which the mobile sRNA loci identified overlap in the three pairwise comparisons. Of 20,679 loci identified from all comparisons, 7179 were associated with mobility and are represented here. (D) Genome features associated with total sRNA, mobile sRNA, and non-differentially represented sRNA loci from the *Col-dcl2,3,4/C24* versus *Col/C24* pairwise comparison were assigned on the basis of annotated features of the *Arabidopsis* genome. Asterisks mark differences between total sRNA and mobile sRNA loci that are statistically significant (Fisher's exact test, $P < 10^{-8}$).

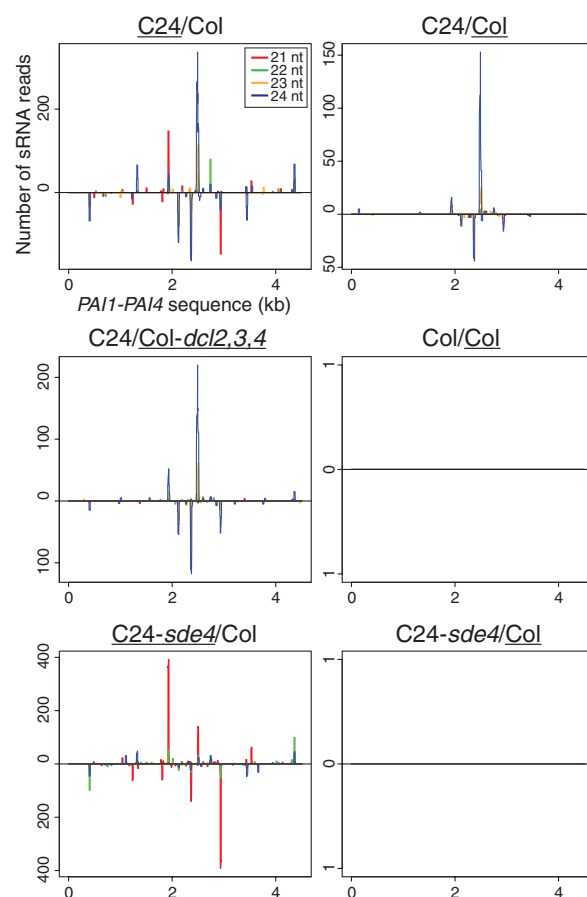
produced by the sequential action of Pol IV and DCL3 in the source tissue, because there were very few 23- and 24-nt sRNAs in C24-*sde4/Col-dcl2,3,4* roots (Fig. 2A and table S2A). Moreover, levels of 23- and 24-nt sRNAs in grafted roots of wild-type Col or C24 plants were lower if they were grafted to Col-*dcl2,3,4* or C24-*sde4* shoots rather than wild-type shoots (fig. S2, A and B).

The mobile 23- and 24-nt sRNA populations are likely to be derived from many different genomic loci because they include a high proportion of the nonredundant sequence reads (Fig. 2A, table S2B). This interpretation is reinforced by pairwise comparisons based on the representation of genomic sRNA in data sets from grafted roots (table S3). In each pair the shoots differed in that only one was competent to generate 23- and 24-nt sRNAs that would migrate into the root. Thus, the mobile sRNA loci would be more abundantly represented in C24/Col-*dcl2,3,4* than in Col-*dcl2,3,4/Col-dcl2,3,4* data sets; the same would be true for Col/C24 (relative to Col-*dcl2,3,4/C24*) and for C24/Col (relative to C24-*sde4/Col*) sample data sets (Fig. 2B and fig. S3, A and B). The degree of differential representation would be high if all of the locus-specific sRNA were mobile, and would be correspondingly lower if the receiving tissue contained a mixture of RNAs that had either moved from the shoot or were produced in situ. Our analysis focused on loci with more than 100 sRNA reads in at least one data set and with a >90% chance of differential representation based on the analysis of replicates.

In these pairwise comparisons the observed differential locus representation was predominantly by a factor of 2 to 11, which is greater than the difference between biological replicates (Fig. 2B and fig. S3A). Of the 20,679 total sRNA loci, 7179 loci were identified in one or more of the comparisons as producing mobile sRNA (Fig. 2C). In a parallel analysis considering the sRNAs only if they matched a unique site in the Col genome, 795 of 2900 sRNA loci were differentially represented (fig. S3, B and C). It is therefore likely that a high proportion of the genomic sRNA loci produce mobile 23- and 24-nt sRNAs, with a substantial proportion being loci from which most of the sRNAs migrate from the shoot to the root. The sRNA data from the 100 loci with highest posterior likelihoods of producing mobile sRNAs identified in each of these pairwise comparisons are shown in tables S5 to S7.

Although most of the mobile sRNA loci did not correspond to genes (Fig. 2D), a notable exception involved the phosphoribosylanthranilate isomerase (*PAI*) tryptophan biosynthetic genes of the C24 genotype. C24 has a tail-to-tail inverted repeat (*PAII-PAI4*) that substitutes for a single copy of *PAI* in Col (fig. S4A) (17). The inverted-repeat DNA has dense cytosine methylation and might silence the other unlinked C24 *PAI* loci through sRNA-directed DNA methylation, although the predicted *PAI* sRNAs have not previously been detected (17, 18).

Fig. 3. Movement of *PAI* inverted-repeat sRNAs. sRNA sequences from grafted *Arabidopsis* were aligned at the first nucleotide position to the *PAI1-PAI4* coding sequence; the positive or negative y axis shows the number of reads at each position on either plus or minus strand. See fig. S4 for an independent biological replicate, and table S1 for a summary of sRNA library details.



We detected 21- to 24-nt sRNAs originating from the C24 *PAI1-PAI4* locus that are absent from the Col data sets (Fig. 3 and fig. S4B). The 24-nt *PAI* siRNAs are present in the C24/Col and C24/Col-*dcl2,3,4* data sets, but, consistent with their mobility between the shoot and root, they are absent in C24-*sde4/Col*. This analysis of the *PAI* locus additionally illustrates a preference for mobility of 24-nt rather than 21-nt sRNAs at this locus. In the shoot samples of a C24-*sde4* plant, the *PAI* 24-nt sRNAs were much less abundant than in the wild-type C24 (Fig. 3 and fig. S4B), but the 21-nt species that do not require Pol IV were unaffected or slightly increased. However, these 21-nt *PAI* sRNAs were almost absent from a C24-*sde4/Col* data set, which indicates that (unlike the 24-nt siRNAs) they are not mobile. This analysis therefore suggests that the bias toward 24-nt species in the mobile sRNAs is not only because they are more numerous than the 21-nt species: There might also be a link between the Pol IV pathway for biogenesis of the 24-nt sRNAs and the transport pathway that mobilizes sRNA between cells and in the phloem.

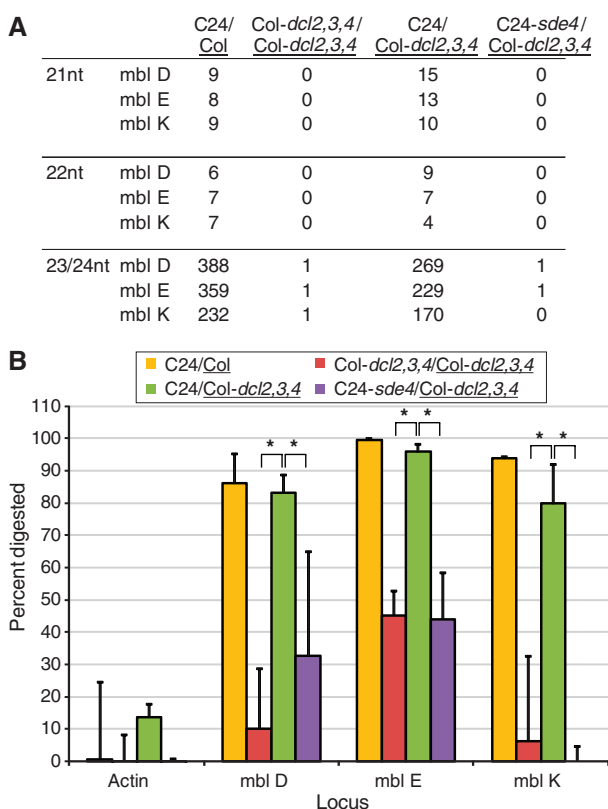
Mobile sRNAs direct DNA methylation. In Fig. 2D we show that the mobile sRNA loci are typical of genomic loci producing 24-nt sRNAs (16). Genes and pseudogenes are associated with only a minority of these loci; the majority are associated with transposons and methylated

DNA in genome-wide analyses of *Arabidopsis* chromatin (19). Prompted by the particularly strong correlation with methylated DNA, we addressed the possibility that mobile sRNA could direct DNA methylation in the recipient tissue.

We first looked for genomic regions that produce mobile 23- and 24-nt sRNAs and at which DNA methylation in roots is reduced relative to the wild type in a Col-*dcl2,3,4* mutant. We selected eight loci with Col-*DCL2,3,4*-dependent or C24-specific DNA methylation in nongrafted samples that was detectable by digestion of genomic DNA with the methylation-dependent enzyme Msp BC (fig. S5A). At three of these loci (referred to as mbl D, E, and K), there was hypermethylation in the C24/Col-*dcl2,3,4* root relative to Col-*dcl2,3,4/Col-dcl2,3,4*. This result is consistent with a sRNA signal from the shoot moving into the root and guiding DNA methylation (Fig. 4 and fig. S5B). When the silencing signal source had an *SDE4* mutation, the production of 24-nt sRNAs at loci mbl D, E, and K was suppressed (Fig. 4A), and, when grafted to a Col-*dcl2,3,4* root, it was unable to rescue DNA methylation at these loci (Fig. 4B and fig. S5B). It is therefore likely that the mobile sRNA responsible for DNA methylation at these three loci is the 24-nt size class produced by Pol IV.

Loci mbl D and E corresponded to methylated upstream or promoter regions of MuDR-type transposable elements (AT3TE73255 and

Fig. 4. Endogenous mobile sRNAs direct root DNA methylation. **(A)** sRNA sequence reads corresponding to three mobile sRNA loci that associate with methylated genomic DNA. The sRNA read number for each data set was normalized to 1,000,000 genome-matching sRNA reads. **(B)** DNA digestion by the DNA methylation-dependent enzyme Mcr BC was used to assess the degree of methylation at mobile loci mbl D, E, and K in samples from grafted plants. Asterisks show statistically significant ($P < 0.01$, Student's t test) differences in root DNA methylation that can be attributed to DCL2, 3, 4 or SDE4 function in the shoot. Error bars indicate 95% confidence intervals from two to four biological replicates.



AT2TE04060, respectively), whereas locus mbl K was in a methylated region downstream of a Harbinger-type transposable element (AT5TE00870). The effect of the mobile sRNA on DNA methylation was greater with loci mbl D and K than with locus mbl E (Fig. 4B and fig. S5B). This difference could be because loci mbl D and K have predominantly asymmetric C methylation (CHH) observed in epigenome maps (20) that is a diagnostic feature of RNA-directed DNA methylation, whereas locus mbl E has a higher level of symmetric CG methylation that might be independent of sRNA. Bisulfite sequencing of locus mbl E confirmed a partial restoration of CHH and CHG methylation (fig. S6).

Discussion. Our data clearly show that DCL-generated 24-nt sRNAs are mobile in plants. It is likely that these mobile 24-nt sRNAs are produced by DCL3 (15) and that they are a subset of the 24-nt sRNAs that have been associated with epigenetic effects (12, 16, 21, 22). Long-range mobility of 24-nt RNAs is consistent with findings that sRNA of this size is present in phloem sap samples (23) and also with molecular analyses showing a correlation between the presence of 24-nt sRNA and systemic silencing in *Nicotiana benthamiana* (24). However, movement of sRNA is not an exclusive property of the 23- and 24-nt sRNA size class; we also found that 22-nt sRNAs produced by DCL2 (15, 25) are mobile (Fig. 1D and fig. S1G). Unfortunately, we cannot use our approach to assay the mobility of 21-nt sRNAs because of the unavailability of viable mutants that completely suppress 21-nt

sRNAs. However, given the presence of 21-nt GFP sRNAs in the transgene grafting experiments (Fig. 1D and figs. S1G and S7) and other recent data, we consider it likely that there are mobile sRNAs in all size classes. Mobility is influenced by features of the genomic locus, including the cell type in which the sRNAs accumulate, in addition to features of individual sRNA. The *PAI* locus will help to unravel these complexities because it produces mobile 23- and 24-nt species and nonmobile 21- and 22-nt sRNAs.

In proposing biological roles for mobile 24-nt sRNAs, we have taken into account the possibility that they move into the shoot and to the shoot meristem, where they may introduce epigenetic effects (Fig. 1 and fig. S7) (1, 11). Our finding that very rare sRNAs such as those in the transgene assay (Fig. 1 and fig. S1) can have an effect may also be relevant, because it implies the existence of an amplification mechanism that may also affect endogenous sRNAs. On the basis of these considerations, we propose two possible roles of the mobile 24-nt sRNAs moving into the shoot meristem. These RNAs could allow the meristem to respond to even transient activation of transposons in somatic cells and thereby reinforce epigenetic silencing of these potentially damaging genetic elements in the cells that give rise to the next generation. It could also be that the mobile sRNAs mediate responses to external stimuli and that they initiate epigenetic changes in the meristem that influence competency to flower or adaptation to stress (26, 27). In the developing seed or pollen (28, 29), the

mobile sRNAs could induce transgenerational epigenetic changes that adapt progeny to future stress.

References and Notes

- O. Voinnet, D. C. Baulcombe, *Nature* **389**, 553 (1997).
- O. Voinnet, P. Vain, S. Angell, D. C. Baulcombe, *Cell* **95**, 177 (1998).
- J. C. Palauqui, T. Elmayan, J. M. Pollien, H. Vaucheret, *EMBO J.* **16**, 4738 (1997).
- P. Dunoyer, C. Himber, O. Voinnet, *Nat. Genet.* **37**, 1356 (2005).
- B. D. Pant, A. Buhtz, J. Kehr, W. R. Scheible, *Plant J.* **53**, 731 (2008).
- C. A. Brosnan *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 14741 (2007).
- T. Dalmay, A. J. Hamilton, E. Mueller, D. C. Baulcombe, *Plant Cell* **12**, 369 (2000).
- S. M. Leisner, R. Turgeon, S. H. Howell, *Plant Cell* **5**, 191 (1993).
- T. J. Lough, W. J. Lucas, *Annu. Rev. Plant Biol.* **57**, 203 (2006).
- I. R. Henderson *et al.*, *Nat. Genet.* **38**, 721 (2006).
- F. Schwach, F. E. Vaistij, L. Jones, D. C. Baulcombe, *Plant Physiol.* **138**, 1842 (2005).
- A. J. Herr, M. B. Jensen, T. Dalmay, D. C. Baulcombe, *Science* **308**, 118 (2005); published online 3 February 2005 (10.1126/science.1106910).
- Arabidopsis Genome Initiative, *Nature* **408**, 796 (2000).
- M. Nordborg *et al.*, *PLoS Biol.* **3**, e196 (2005).
- Z. Xie *et al.*, *PLoS Biol.* **2**, e104 (2004).
- R. A. Mosher, F. Schwach, D. Studholme, D. C. Baulcombe, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 3145 (2008).
- S. Melquist, B. Luff, J. Bender, *Genetics* **153**, 401 (1999).
- S. Melquist, J. Bender, *Genes Dev.* **17**, 2036 (2003).
- X. Zhang *et al.*, *Cell* **126**, 1189 (2006).
- R. Lister *et al.*, *Cell* **133**, 523 (2008).
- Y. Onodera *et al.*, *Cell* **120**, 613 (2005).
- X. Zhang, I. R. Henderson, C. Lu, P. J. Green, S. E. Jacobsen, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 4536 (2007).
- A. Buhtz, F. Springer, L. Chappell, D. C. Baulcombe, J. Kehr, *Plant J.* **53**, 739 (2008).
- A. J. Hamilton, O. Voinnet, L. Chappell, D. C. Baulcombe, *EMBO J.* **21**, 4671 (2002).
- A. Deleris *et al.*, *Science* **313**, 68 (2006); published online 1 June 2006 (10.1126/science.1128214).
- I. Bäurle, L. Smith, D. C. Baulcombe, C. Dean, *Science* **318**, 109 (2007).
- O. Borsani, J. Zhu, P. E. Verslues, R. Sunkar, J.-K. Zhu, *Cell* **123**, 1279 (2005).
- R. A. Mosher *et al.*, *Nature* **460**, 283 (2009).
- R. K. Slotkin *et al.*, *Cell* **136**, 461 (2009).
- We thank K. Rutherford for data assembly and the Genomics and Bioinformatics core facilities (CRUK Cambridge Research Institute) for Illumina sequencing. Supported by Commonwealth scholarship CACR 2006-115 and Natural Sciences and Engineering Research Council of Canada scholarship PGS D2 358506-2008 (C.W.M.), a Royal Society research professorship (D.C.B.), the Gatsby Charitable Foundation, the European Union Sixth Framework Programme Integrated Project SIROCCO (grant LSHG-CT-2006-037900), and European Research Council advanced investigator grant 233325 REVOLUTION. RNA sequencing raw data are available from Gene Expression Omnibus (GSE20663).

Supporting Online Material

www.sciencemag.org/cgi/content/full/science.1187959/DC1
Materials and Methods
Figs. S1 to S7
Tables S1 to S7
References

4 February 2010; accepted 26 March 2010
Published online 22 April 2010;
10.1126/science.1187959
Include this information when citing this paper.

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

The following resources related to this article are available online at www.sciencemag.org (this information is current as of May 13, 2010):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/328/5980/876>

Supporting Online Material can be found at:

<http://www.sciencemag.org/cgi/content/full/328/5980/876/DC1>

This article **cites 19 articles**, 3 of which can be accessed for free:

<http://www.sciencemag.org/cgi/content/full/328/5980/876#otherarticles>

This article appears in the following **subject collections**:

Computers, Mathematics

http://www.sciencemag.org/cgi/collection/comp_math

Community Structure in Time-Dependent, Multiscale, and Multiplex Networks

Peter J. Mucha,^{1,2*} Thomas Richardson,^{1,3} Kevin Macon,¹ Mason A. Porter,^{4,5} Jukka-Pekka Onnela^{6,7}

Network science is an interdisciplinary endeavor, with methods and applications drawn from across the natural, social, and information sciences. A prominent problem in network science is the algorithmic detection of tightly connected groups of nodes known as communities. We developed a generalized framework of network quality functions that allowed us to study the community structure of arbitrary multislice networks, which are combinations of individual networks coupled through links that connect each node in one network slice to itself in other slices. This framework allows studies of community structure in a general setting encompassing networks that evolve over time, have multiple types of links (multiplexity), and have multiple scales.

The study of graphs, or networks, has a long tradition in fields such as sociology and mathematics, and it is now ubiquitous in academic and everyday settings. An important tool in network analysis is the detection of mesoscopic structures known as communities (or cohesive groups), which are defined intuitively as groups of nodes that are more tightly connected to each other than they are to the rest of the network (1–3). One way to quantify communities is by a quality function that compares the number of intracommunity edges to what one would expect at random. Given the network adjacency matrix A , where the element A_{ij} details a direct connection between nodes i and j , one can construct a quality function $Q(4, 5)$ for the partitioning of nodes into communities as $Q = \sum_{ij} (A_{ij} - P_{ij}) \delta(g_i, g_j)$, where $\delta(g_i, g_j) = 1$ if the community assignments g_i and g_j of nodes i and j are the same and 0 otherwise, and P_{ij} is the expected weight of the edge between i and j under a specified null model.

The choice of null model is a crucial consideration in studying network community structure (2). After selecting a null model appropriate to the network and application at hand, one can use a variety of computational heuristics to assign nodes to communities to optimize the quality Q (2, 3). However, such null models have not been available for time-dependent networks; analyses have instead depended on ad hoc methods to

piece together the structures obtained at different times (6–9) or have abandoned quality functions in favor of such alternatives as the Minimum Description Length principle (10). Although tensor decompositions (11) have been used to cluster network data with different types of connections, no quality-function method has been developed for such multiplex networks.

We developed a methodology to remove these limits, generalizing the determination of community structure via quality functions to multislice networks that are defined by coupling multiple adjacency matrices (Fig. 1). The connections encoded by the network slices are flexible; they can represent variations across time, variations across different types of connections, or even community detection of the same network at different scales. However, the usual procedure for establishing a quality function as a direct count of the intracommunity edge weight minus that

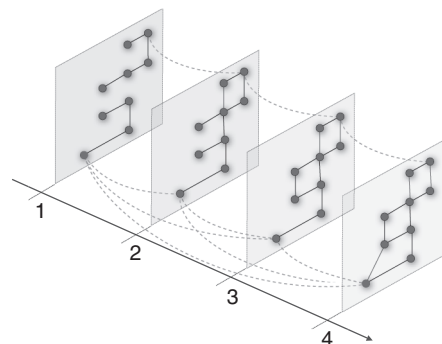


Fig. 1. Schematic of a multislice network. Four slices $s = \{1, 2, 3, 4\}$ represented by adjacencies A_{js} encode intraslice connections (solid lines). Interslice connections (dashed lines) are encoded by C_{jrs} , specifying the coupling of node j to itself between slices r and s . For clarity, interslice couplings are shown for only two nodes and depict two different types of couplings: (i) coupling between neighboring slices, appropriate for ordered slices; and (ii) all-to-all interslice coupling, appropriate for categorical slices.

expected at random fails to provide any contribution from these interslice couplings. Because they are specified by common identifications of nodes across slices, interslice couplings are either present or absent by definition, so when they do fall inside communities, their contribution in the count of intra-community edges exactly cancels that expected at random. In contrast, by formulating a null model in terms of stability of communities under Laplacian dynamics, we have derived a principled generalization of community detection to multislice networks,

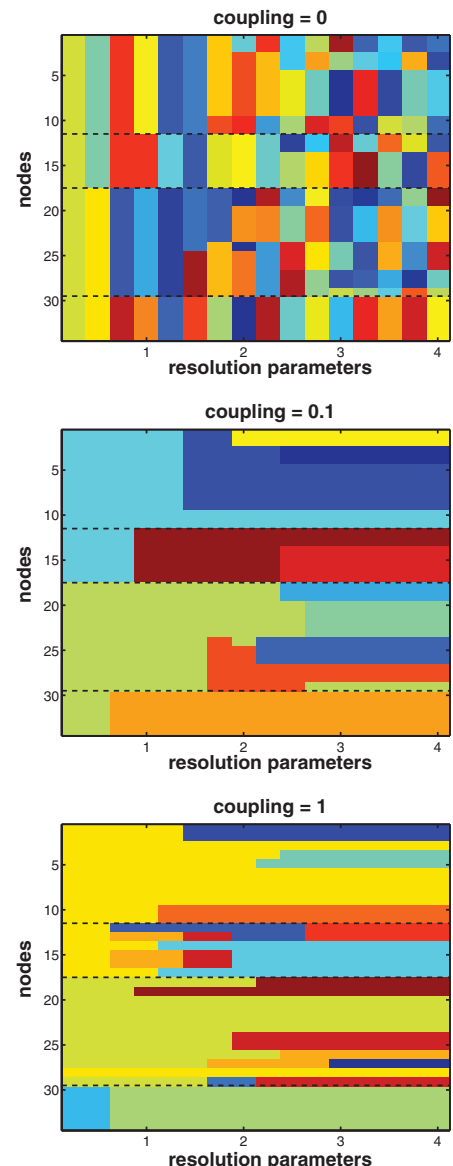


Fig. 2. Multislice community detection of the Zachary Karate Club network (22) across multiple resolutions. Colors depict community assignments of the 34 nodes (renumbered vertically to group similarly assigned nodes) in each of the 16 slices (with resolution parameters $\gamma_s = \{0.25, 0.5, \dots, 4\}$), for $\omega = 0$ (top), $\omega = 0.1$ (middle), and $\omega = 1$ (bottom). Dashed lines bound the communities obtained using the default resolution ($\gamma = 1$).

¹Carolina Center for Interdisciplinary Applied Mathematics, Department of Mathematics, University of North Carolina, Chapel Hill, NC 27599, USA. ²Institute for Advanced Materials, Nanoscience and Technology, University of North Carolina, Chapel Hill, NC 27599, USA. ³Operations Research, North Carolina State University, Raleigh, NC 27695, USA. ⁴Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, University of Oxford, Oxford OX1 3LB, UK. ⁵CABDyN Complexity Centre, University of Oxford, Oxford OX1 1HP, UK. ⁶Department of Health Care Policy, Harvard Medical School, Boston, MA 02115, USA. ⁷Harvard Kennedy School, Harvard University, Cambridge, MA 02138, USA.

*To whom correspondence should be addressed. E-mail: mucha@unc.edu

with a single parameter controlling the interslice correspondence of communities.

Important to our method is the equivalence between the modularity quality function (12) [with a resolution parameter (5)] and stability of communities under Laplacian dynamics (13), which we have generalized to recover the null models for bipartite, directed, and signed networks (14). First, we obtained the resolution-parameter generaliza-

tion of Barber's null model for bipartite networks (15) by requiring the independent joint probability contribution to stability in (13) to be conditional on the type of connection necessary to step between two nodes. Second, we recovered the standard null model for directed networks (16, 17) (again with a resolution parameter) by generalizing the Laplacian dynamics to include motion along different kinds of connections—in this case,

both with and against the direction of a link. By this generalization, we similarly recovered a null model for signed networks (18). Third, we interpreted the stability under Laplacian dynamics flexibly to permit different spreading weights on the different types of links, giving multiple resolution parameters to recover a general null model for signed networks (19).

We applied these generalizations to derive null models for multislice networks that extend the existing quality-function methodology, including an additional parameter ω to control the coupling between slices. Representing each network slice s by adjacencies A_{ijs} between nodes i and j , with interslice couplings C_{jrs} that connect node j in slice r to itself in slice s (Fig. 1), we have restricted our attention to unipartite, undirected network slices ($A_{ijs} = A_{jis}$) and couplings ($C_{jrs} = C_{jsr}$), but we can incorporate additional structure in the slices and couplings in the same manner as demonstrated for single-slice null models. Notating the strengths of each node individually in each slice by $k_{js} = \sum_i A_{ijs}$ and across slices by $c_{js} = \sum_r C_{jrs}$, we define the multislice strength by $\kappa_{js} = k_{js} + c_{js}$. The continuous-time Laplacian dynamics given by

$$\dot{p}_{is} = \sum_{jr} \frac{(A_{ijs} \delta_{sr} + \delta_{ij} C_{jsr}) p_{jr}}{\kappa_{jr}} - p_{is} \quad (1)$$

respects the intraslice nature of A_{ijs} and the interslice couplings of C_{jsr} . Using the steady-state probability distribution $p_{jr}^* = \kappa_{jr}/2\mu$, where $2\mu = \sum_{jr} \kappa_{jr}$, we obtained the multislice null model in terms of the probability $\rho_{is|jr}$ of sampling node i in slice s conditional on whether the multislice structure allows one to step from (j, r) to (i, s) , accounting for intra- and interslice steps separately as

$$\rho_{is|jr} p_{jr}^* = \left(\frac{k_{is}}{2m_s} \frac{k_{jr}}{\kappa_{jr}} \delta_{sr} + \frac{C_{jsr} c_{jr}}{c_{jr} \kappa_{jr}} \delta_{ij} \right) \frac{\kappa_{jr}}{2\mu} \quad (2)$$

where $m_s = \sum_j k_{js}$. The second term in parentheses, which describes the conditional probability of motion between two slices, leverages the definition of the C_{jsr} coupling. That is, the conditional probability of stepping from (j, r) to (i, s) along an interslice coupling is nonzero if and only if $i = j$, and it is proportional to the probability C_{jsr}/κ_{jr} of selecting the precise interslice link that connects to slice s . Subtracting this conditional joint probability from the linear (in time) approximation of the exponential describing the Laplacian dynamics, we obtained a multislice generalization of modularity (14):

$$\mathcal{Q}_{\text{multislice}} = \frac{1}{2\mu} \sum_{ijsr} \left[\left(A_{ijs} - \gamma_s \frac{k_{is} k_{js}}{2m_s} \delta_{sr} \right) + \delta_{ij} C_{jsr} \right] \delta(g_{is}, g_{jr}) \quad (3)$$

where we have used reweighting of the conditional probabilities, which allows a different resolution γ_s in each slice. We have absorbed the resolution parameter for the interslice couplings into the magnitude of the elements of C_{jsr} , which, for simplicity, we presume to take binary values $\{0, \omega\}$ indicating the absence (0) or presence (ω) of interslice links.

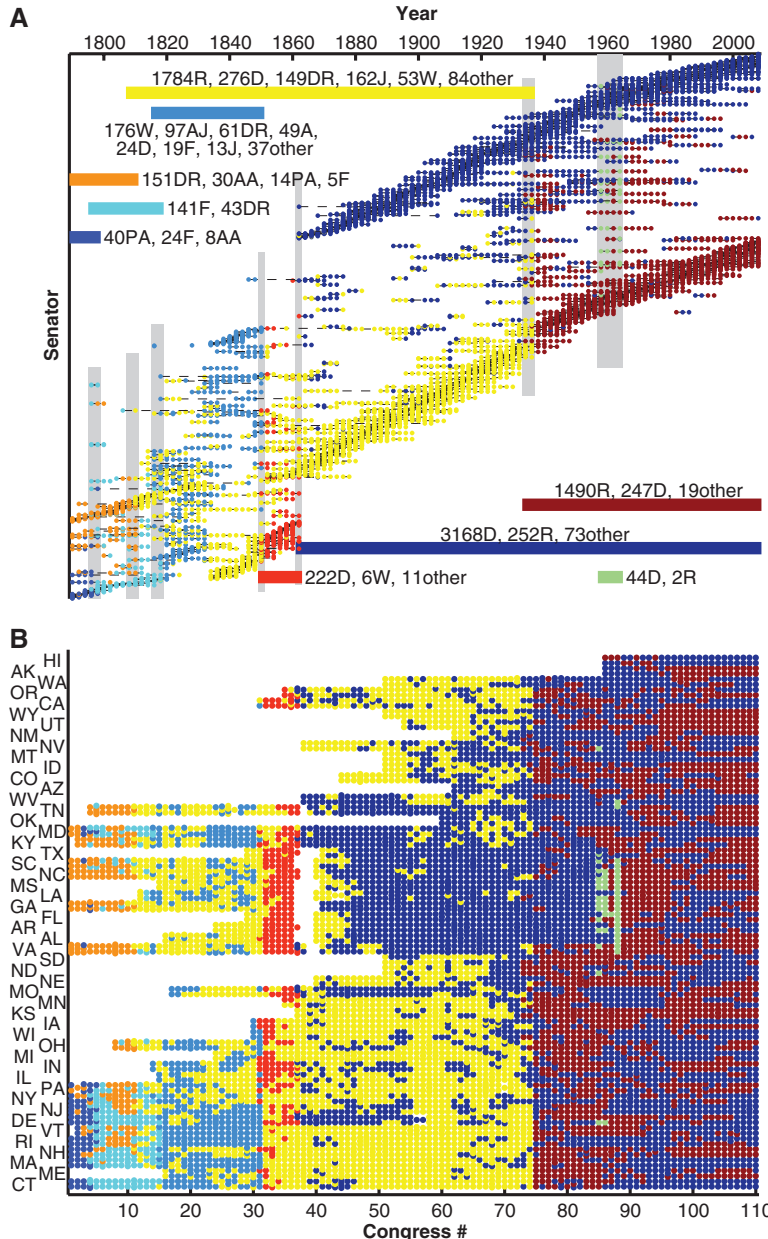


Fig. 3. Multislice community detection of U.S. Senate roll call vote similarities (23) with $\omega = 0.5$ coupling of 110 slices (i.e., the number of 2-year Congresses from 1789 to 2008) across time. **(A)** Colors indicate assignments to nine communities of the 1884 unique senators (sorted vertically and connected across Congresses by dashed lines) in each Congress in which they appear. The dark blue and red communities correspond closely to the modern Democratic and Republican parties, respectively. Horizontal bars indicate the historical period of each community, with accompanying text enumerating nominal party affiliations of the single-slice nodes (each representing a senator in a Congress): PA, pro-administration; AA, anti-administration; F, Federalist; DR, Democratic-Republican; W, Whig; AJ, anti-Jackson; A, Adams; J, Jackson; D, Democratic; R, Republican. Vertical gray bars indicate Congresses in which three communities appeared simultaneously. **(B)** The same assignments according to state affiliations.

Table 1. Communities in the first wave of the multiplex “Tastes, Ties, and Time” network (24), using the default resolution ($\gamma = 1$) in each of the four slices of data (Facebook friendships, picture friendships, roommates, and housing groups) under various couplings ω across slices, which changed the number of communities and percentages of individuals assigned on a per-slice basis to one, two, three, or four communities.

ω	Number of communities	Communities per individual (%)			
		1	2	3	4
0	1036	0	0	0	100
0.1	122	14.0	40.5	37.3	8.2
0.2	66	19.9	49.1	25.3	5.7
0.3	49	26.2	48.3	21.6	3.9
0.4	36	31.8	47.0	18.4	2.8
0.5	31	39.3	42.4	16.8	1.5
1	16	100	0	0	0

Community detection in multislice networks can then proceed using many of the same computational heuristics that are currently available for single-slice networks [although, as with the standard definition of modularity, one must be cautious about the resolution of communities (20) and the likelihood of complex quality landscapes that necessitate caution in interpreting results on real networks (21)]. We studied examples that have multiple resolutions [Zachary Karate Club (22)], vary over time [voting similarities in the U.S. Senate (23)], or are multiplex [the “Tastes, Ties, and Time” cohort of university students (24)]. We provide additional details for each example in (14).

We performed simultaneous community detection across multiple resolutions (scales) in the well-known Zachary Karate Club network, which encodes the friendships between 34 members of a 1970s university karate club (22). Keeping the same unweighted adjacency matrix across slices ($A_{ijs} = A_{ij}$ for all s), the resolution associated with each slice is dictated by a specified sequence of γ_s parameters, which we chose to be the 16 values $\gamma_s = \{0.25, 0.5, 0.75, \dots, 4\}$. In Fig. 2, we depict the community assignments obtained for coupling strengths $\omega = \{0, 0.1, 1\}$ between each neighboring pair of the 16 ordered slices. These results simultaneously probe all scales, including the partition of the Karate Club into four communities at the default resolution of modularity (3, 25). Additionally, we identified nodes that have an especially strong tendency to break off from larger communities (e.g., nodes 24 to 29 in Fig. 2).

We also considered roll call voting in the U.S. Senate across time, from the 1st Congress to the 110th, covering the years 1789 to 2008 and including 1884 distinct senator IDs (26). We defined weighted connections between each pair of senators by a similarity between their voting, specified independently for each 2-year Congress (23). We studied the multislice collection of these 110 networks, with each individual senator coupled to himself or herself when appearing in consecutive Congresses. Multislice community detection uncovered interesting details about the continuity of individual and group voting trends over time that are not captured by the union of the 110 independent partitions of the separate Congresses.

Figure 3 depicts a partition into nine communities that we obtained using coupling $\omega = 0.5$. The Congresses in which three communities appeared simultaneously are each historically noteworthy: The 4th and 5th Congresses were the first with political parties; the 10th and 11th Congresses occurred during the political drama of former Vice President Aaron Burr’s indictment for treason; the 14th and 15th Congresses witnessed the beginning of changing group structures in the Democratic-Republican party amidst the dying Federalist party (23); the 31st Congress included the Compromise of 1850; the 37th Congress occurred during the beginning of the American Civil War; the 73rd and 74th Congresses followed the landslide 1932 election (during the Great Depression); and the 85th to 88th Congresses brought the major American civil rights acts, including the congressional fights over the Civil Rights Acts of 1957, 1960, and 1964.

Finally, we applied multislice community detection to a multiplex network of 1640 college students at a northeastern American university (24), including symmetrized connections from the first wave of this data representing (i) Facebook friendships, (ii) picture friendships, (iii) roommates, and (iv) student housing-group preferences. Because the different connection types are categorical, the natural interslice couplings connect an individual in a slice to himself or herself in each of the other three network slices. This coupling between categorical slices thus differs from that above, which connected only neighboring (ordered) slices. Table 1 indicates the numbers of communities and the percentages of individuals assigned to one, two, three, or four communities across the four types of connections for different values of ω , as a first investigation of the relative redundancy across the connection types.

Our multislice framework makes it possible to study community structure in a much broader class of networks than was previously possible. Instead of detecting communities in one static network at a time, our formulation generalizing the Laplacian dynamics approach of (13) permits the simultaneous quality-function study of community structure across multiple times, multiple resolution parameter values, and multiple types of links. We

used this method to demonstrate insights in real-world networks that would have been difficult or impossible to obtain without the simultaneous consideration of multiple network slices. Although our examples included only one kind of variation at a time, our framework applies equally well to networks that have multiple such features (e.g., time-dependent multiplex networks). We expect multislice community detection to become a powerful tool for studying such systems.

References and Notes

1. M. Girvan, M. E. J. Newman, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 7821 (2002).
2. M. A. Porter, J.-P. Onnela, P. J. Mucha, *Not. Am. Math. Soc.* **56**, 1082 (2009).
3. S. Fortunato, *Phys. Rep.* **486**, 75 (2010).
4. M. E. J. Newman, *Phys. Rev. E* **74**, 036104 (2006).
5. J. Reichardt, S. Bornholdt, *Phys. Rev. E* **74**, 016110 (2006).
6. J. Hopcroft, O. Khan, B. Kulis, B. Selman, *Proc. Natl. Acad. Sci. U.S.A.* **101** (suppl. 1), 5249 (2004).
7. T. Y. Berger-Wolf, J. Saia, in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2006), p. 523 (10.1145/1150402.1150462).
8. G. Palla, A.-L. Barabási, T. Vicsek, *Nature* **446**, 664 (2007).
9. D. J. Fenn et al., *Chaos* **19**, 033119 (2009).
10. J. Sun, C. Faloutsos, S. Papadimitriou, P. S. Yu, in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2007), p. 687 (10.1145/1281192.1281266).
11. T. M. Selee, T. G. Kolda, W. P. Kegelmeyer, J. D. Griffin, *CSRI Summer Proceedings 2007, Technical Report SAND2007-7977*, Sandia National Laboratories, Albuquerque, NM and Livermore, CA, M. L. Parks, S. S. Collis, Eds. (2007), p. 87 (www.cs.sandia.gov/CSRI/Proceedings).
12. M. E. J. Newman, M. Girvan, *Phys. Rev. E* **69**, 026113 (2004).
13. R. Lambiotte, J. C. Delvenne, M. Barahona, <http://arxiv.org/abs/0812.1770> (2008).
14. See supporting material on Science Online.
15. M. J. Barber, *Phys. Rev. E* **76**, 066102 (2007).
16. A. Arenas, J. Duch, A. Fernandez, S. Gomez, *N. J. Phys.* **9**, 176 (2007).
17. E. A. Leicht, M. E. J. Newman, *Phys. Rev. Lett.* **100**, 118703 (2008).
18. S. Gómez, P. Jensen, A. Arenas, *Phys. Rev. E* **80**, 016114 (2009).
19. V. A. Traag, J. Bruggeman, *Phys. Rev. E* **80**, 036115 (2009).
20. S. Fortunato, M. Barthélemy *Proc. Natl. Acad. Sci. U.S.A.* **104**, 36 (2007).
21. B. H. Good, Y.-A. de Montjoye, A. Clauset, *Phys. Rev. E* **81**, 046106 (2010).
22. W. W. Zachary, *J. Anthropol. Res.* **33**, 452 (1977).
23. A. S. Waugh, L. Pei, J. H. Fowler, P. J. Mucha, M. A. Porter, <http://arxiv.org/abs/0907.3509> (2009).
24. K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, N. Christakis, *Soc. Networks* **30**, 330 (2008).
25. T. Richardson, P. J. Mucha, M. A. Porter, *Phys. Rev. E* **80**, 036111 (2009).
26. K. T. Poole, Voteview (<http://voteview.com>) (2008).
27. We thank N. A. Christakis, L. Meneades, and K. Lewis for access to and helping with the “Tastes, Ties, and Time” data; S. Reid and A. L. Traud for help developing code; and A. Clauset, J.-C. Delvenne, S. Fortunato, M. Gould, and V. Traag for discussions. Congressional roll call data are from <http://voteview.com> (26). Supported by NSF grant DMS-0645369 (P.J.M.), James S. McDonnell Foundation grant 220020177 (M.A.P.), and the Fulbright Program (J.-P.O.).

Supporting Online Material

www.sciencemag.org/cgi/content/full/328/5980/876/DC1
SOM Text
References
17 November 2009; accepted 22 March 2010
10.1126/science.1184819

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

The following resources related to this article are available online at www.sciencemag.org (this information is current as of May 13, 2010):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/328/5980/879>

Supporting Online Material can be found at:

<http://www.sciencemag.org/cgi/content/full/328/5980/879/DC1>

This article **cites 29 articles**, 2 of which can be accessed for free:

<http://www.sciencemag.org/cgi/content/full/328/5980/879#otherarticles>

This article has been **cited by 1 articles** hosted by HighWire Press; see:

<http://www.sciencemag.org/cgi/content/full/328/5980/879#otherarticles>

This article appears in the following **subject collections**:

Physics

<http://www.sciencemag.org/cgi/collection/physics>

High-NOON States by Mixing Quantum and Classical Light

Itai Afek, Oron Ambar, Yaron Silberberg*

Precision measurements can be brought to their ultimate limit by harnessing the principles of quantum mechanics. In optics, multiphoton entangled states, known as NOON states, can be used to obtain high-precision phase measurements, becoming more and more advantageous as the number of photons grows. We generated “high-NOON” states ($N = 5$) by multiphoton interference of quantum down-converted light with a classical coherent state in an approach that is inherently scalable. Super-resolving phase measurements with up to five entangled photons were produced with a visibility higher than that obtainable using classical light only.

Entanglement is a distinctive feature of quantum mechanics that lies at the core of many new applications in the emerging science of quantum information. Multiparticle entangled states are central to quantum computing, quantum teleportation, and quantum metrology (1). A particularly useful class of states are the maximally path-entangled multiphoton entangled states (NOON states)

$$|N::0\rangle_{A,B} \equiv \frac{1}{\sqrt{2}} (|N, 0\rangle_{A,B} + |0, N\rangle_{A,B}) \quad (1)$$

Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel.

*To whom correspondence should be addressed. E-mail: yaron.silberberg@weizmann.ac.il

which contain N indistinguishable particles in an equal superposition of all being in one of two possible paths A and B (2). These states are “Schrödinger cats,” as they consist of a superposition of two highly distinct states corresponding to the “dead and alive” cat. The larger the value of N , the bigger the cat is, thus bringing us closer to the regime of macroscopic entanglement envisioned originally by Schrödinger (3). Realization of such states is required for the experimental study of fundamental questions such as the effect of decoherence on many-particle entanglement (4, 5). In addition, NOON states are the enabling technology behind various quantum measurement schemes. In optics, a NOON state with N entangled photons acquires a phase at a rate N times as fast as classical light (6). This

leads to enhanced phase sensitivity, which can be used for reaching the fundamental Heisenberg limit (2), and to phase super-resolution, which is the key to sub-Rayleigh resolution in quantum lithography (7). NOON states based on nuclear spin (8) and atomic spin waves (9) have also been shown to allow enhanced measurement sensitivity.

Our goal is to generate optical NOON states with high photon numbers. Various schemes for generating such states have been suggested (2, 10–15). However, existing realizations (16–18) have been limited to three-photon states. We note that a “NOON-like” four-photon state has been generated, but only by using four rather than two spatial modes (19). In addition, a number of experiments have used state projection to focus on the NOON component of various initial N -photon states (20–23).

We present an experimental realization of an approach that yields NOON states with arbitrarily high photon numbers (24, 25). The underlying principle is that when a classical coherent state and quantum down-converted light are mixed properly using a standard beamsplitter, the emerging state shows “Schrödinger cat”-like behavior; that is, virtually all the photons exit collectively from one of the beamsplitter ports or the other (Fig. 1A). The approach is appealing because of its inherent simplicity, as it relies on a fundamental unmodified multiphoton interference effect. Consider a 50/50 beamsplit-

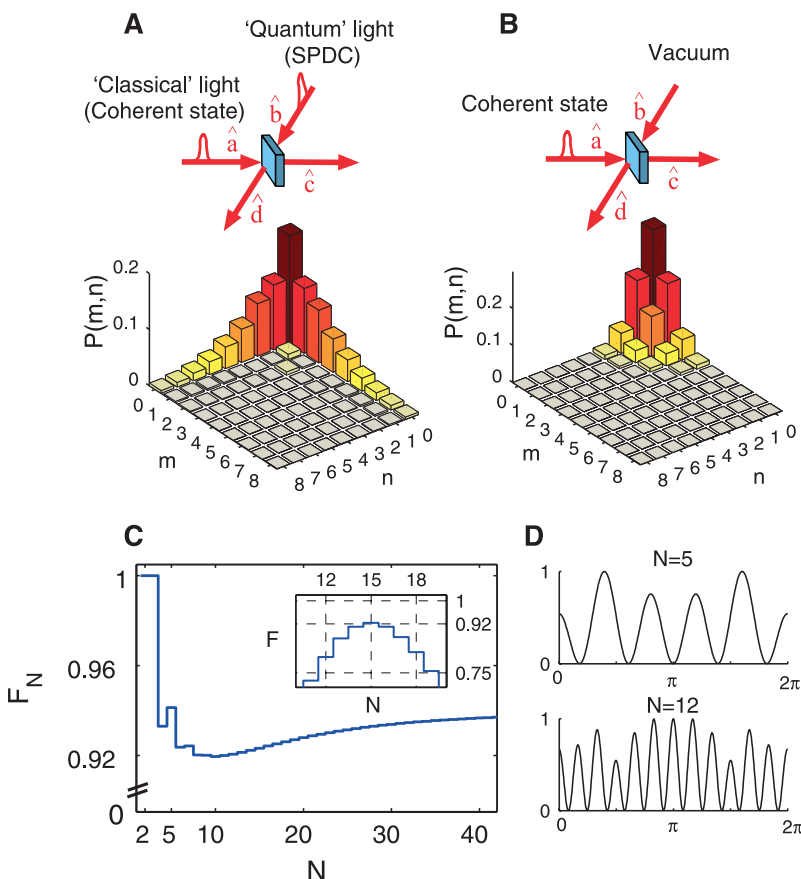


Fig. 1. Theoretical properties of the generated states. (A) Classical light is mixed on a beamsplitter with quantum photon pairs produced by collinear SPDC. Bar heights represent the probability for m, n photons in output modes c, d , respectively. The “corner”-like shape illustrates the tendency of all the photons to collectively exit from the same output port, exhibiting “Schrödinger cat”-like behavior. The effect occurs for arbitrary photon fluxes and is demonstrated here using 1.2 photons per pulse from each input. (B) The same as (A) but using only the classical light, shown for comparison. The photons at the outputs are clearly in a separable (unentangled) state. (C) Fidelity F_N versus N in an ideal setup. The pair amplitude ratio γ , which maximizes the NOON state overlap, was chosen separately for each N . Optimal fidelity is always larger than 0.92, and it approaches 0.943 asymptotically for large N . The inset shows that when γ is optimized for $N = 15$, the fidelity for nearby N is also high. In this case $F > 0.75$ for $N = 12$ to 19, simultaneously. (D) Simulated N -fold coincidences as a function of Mach-Zehnder phase for $N = 5$ or 12, demonstrating N -fold super-resolution.

Fig. 2. Experimental setup. **(A)** Schematic of the setup depicting a Mach-Zehnder interferometer (MZ) fed by a coherent state and SPDC. The NOON states occur in modes $\hat{c}$ and $\hat{d}$ after the first beamsplitter. Measurement of multiphoton coincidences is performed using photon number-resolving detectors. **(B)** Detailed layout of the setup. A pulsed Ti:sapphire oscillator with 120-fs pulses at a repetition rate of 80 MHz is doubled using a 2.74-mm lithium triborate (LBO) crystal to obtain 404-nm ultraviolet pulses with maximum power of 225 mW. These pulses then pump collinear degenerate type I SPDC at 808 nm using a 1.78-mm beta barium borate (BBO) crystal. The SPDC ($\hat{H}$ polarization) is mixed with attenuated coherent light ($\hat{V}$ polarization) using a polarizing beamsplitter (PBS). A thermally induced drift in the relative phase is corrected every few minutes with the use of a liquid crystal (LC) phase retarder. The MZ is polarization-based in a collinear, inherently phase-stable design. The MZ phase is controlled using an additional LC phase retarder at 45° , which adjusts the phase between $\hat{X}$ and $\hat{Y}$ polarizations. The spatial and spectral modes are matched using a polarization-maintaining fiber (PMF) and a 3-nm (full width at half maximum) bandpass filter (BPF). Photon number-resolving detection is performed using an array of single-photon counting modules (SPCM, Perkin Elmer).

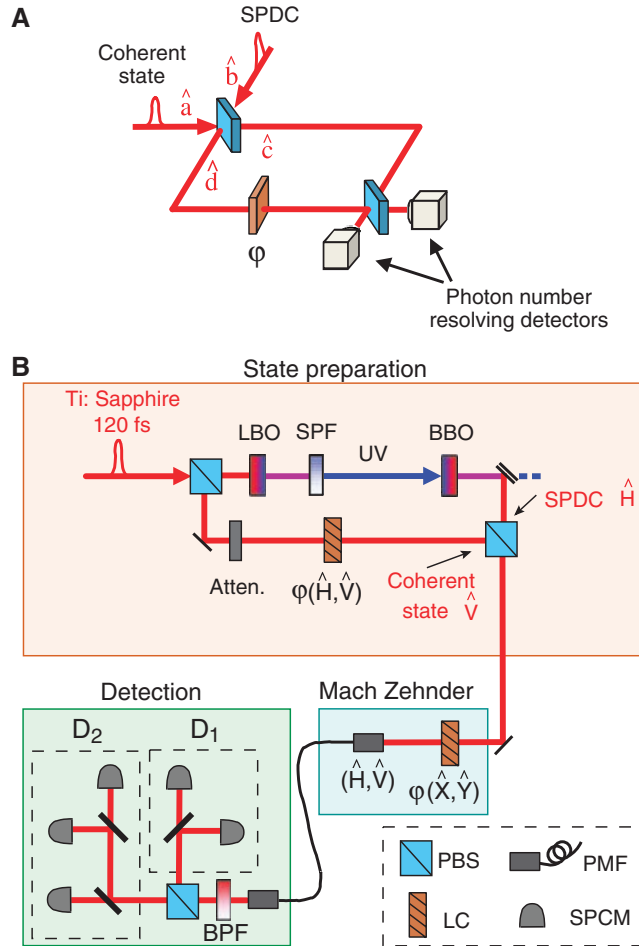
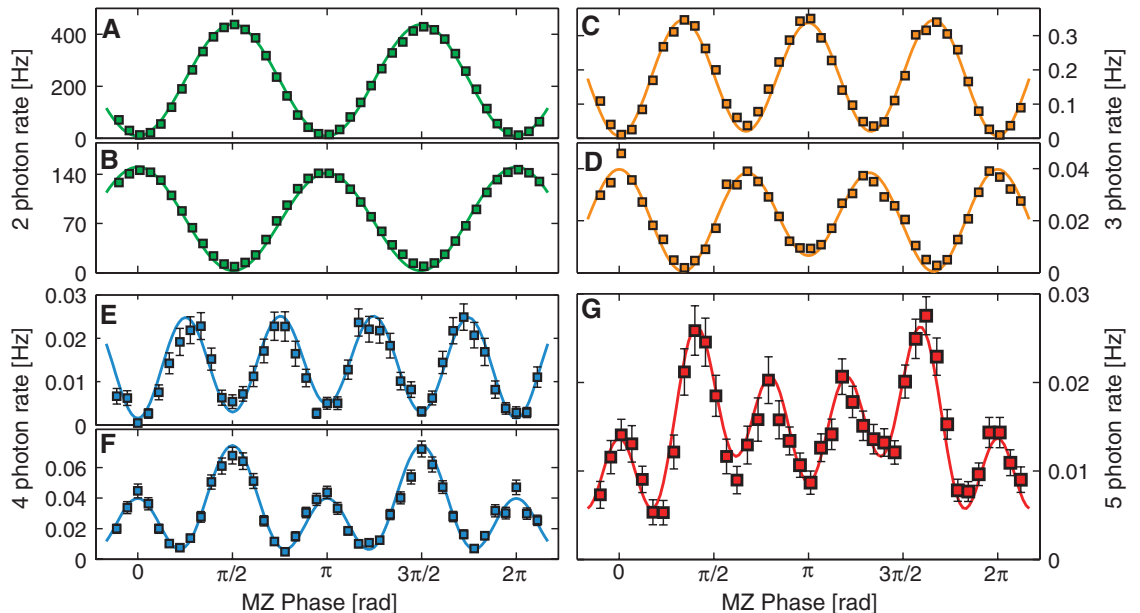


Fig. 3. Experimental results: Coincidence measurements demonstrating N -fold super-resolution for $N = 2, 3, 4$, and 5 with no background subtraction. Error bars indicate $\pm\sigma$ statistical uncertainty. Solid lines are obtained using an analytical model (26). Numbers of simultaneous “clicks” in detectors D_1 and D_2 are denoted N_1 and N_2 , respectively (Fig. 2B). **(A and B)** Two-photon rate with $N_1, N_2 = 1, 1$ (A) and $N_1, N_2 = 2, 0$ (B). **(C and D)** Three-photon rate with $N_1, N_2 = 2, 1$ (C) and $N_1, N_2 = 3, 0$ (D). **(E and F)** Four-photon rate with $N_1, N_2 = 3, 1$ (E) and $N_1, N_2 = 2, 2$ (F). **(G)** Five-photon rate with $N_1, N_2 = 3, 2$. For each N the pair amplitude ratio γ was chosen separately to maximize the NOON state fidelity. The optimal values (obtained analytically) are $\gamma_2 = \gamma_3 = 1$, $\gamma_4 = \sqrt{3}$, and $\gamma_5 = 9/(\sqrt{10} + 1) \approx 2.16$. Visibility of the sinusoidal patterns with N oscillations is determined by



a weighted least-squares sine-cosine decomposition restricted to frequencies of $0, 1, \dots, N$ (27). The values for plots (A) to (G) are $V = 95 \pm 0.0\%$, $88 \pm 0.03\%$, $86 \pm 0.6\%$, $80 \pm 1.9\%$, $74 \pm 3\%$, $73 \pm 2.4\%$, and $42 \pm 2\%$, respectively.

ter fed by a coherent state, $|\alpha\rangle_{cs}$, in one input port and collinear degenerate spontaneous parametric down-conversion (SPDC), $|\xi\rangle_{sb}$, in the other (Fig. 1A). The input states are defined in the conventional way (6):

$$|\alpha\rangle = \sum_{n=0}^{\infty} \exp\left(-\frac{1}{2}|\alpha|^2\right) \frac{\alpha^n}{\sqrt{n!}} |n\rangle, \quad \alpha = |\alpha| \exp(i\theta_{cs}) \quad (2)$$

$$|\xi\rangle = \frac{1}{\sqrt{\cosh r}} \sum_{m=0}^{\infty} (-1)^m \frac{\sqrt{(2m)!}}{2^m m!} (\tanh r)^m |2m\rangle \quad (3)$$

where the phase of $|\xi\rangle$ has been set arbitrarily to zero, leaving the relative phase of the two inputs to be determined by θ_{cs} . We denote the pair amplitude ratio of the coherent state and SPDC inputs by $\gamma \equiv |\alpha|^2/r$. In physical terms, γ^2 is the two-photon probability of the classical source divided by that of the quantum source. The larger the value of γ , the higher the relative weight of the classical resources. The state at the beamsplitter output, $|\psi_{out}\rangle_{c,d}$, is highly path-entangled. A general N -photon two-mode state can be written as $\sum_{k=0}^N u_k |k\rangle_c |N-k\rangle_d$. The creation of an ideal NOON state would require elimination of all the “non-NOON” components (i.e., $u_1, \dots, u_{N-1} = 0$). The present scheme does this almost perfectly by using the naturally emerging multiphoton interference (Fig. 1A). The fidelity of the output state’s normalized N -photon component, $|\psi_{out}^N\rangle$, with a NOON state is $F_N \equiv |\langle N::0 | \psi_{out}^N \rangle|^2$. It can be shown that by choosing $\theta_{cs} = \pi/2$ and optimizing γ for each N , one can achieve $F_N > 0.92$

for arbitrary N (Fig. 1C). The theoretical overlaps for the states generated in this work are $F_N = 1, 1, 0.933$, and 0.941 for $N = 2, 3, 4$, and 5 , respectively. We note that the four-photon value is much higher than the theoretical fidelity of 0.75 obtainable using down-conversion only (22, 23).

To verify the N -photon coherence of the generated states, we use a Mach-Zehnder interferometer (Fig. 2). This is the prototypical setup used in schemes for quantum lithography and reduced-noise interferometry (2). The quantum and classical inputs are prepared in perpendicular linear polarizations (H, V) and spatially overlapped using a polarizing beamsplitter. A Mach-Zehnder interferometer (MZ) is then implemented in an inherently phase-stable, collinear design so that the NOON state mode subscripts c, d may now be replaced by X, Y ($\pm 45^\circ$ polarizations). After application of the MZ phase shift ϕ , the state $|N, 0\rangle_{XY} + |0, N\rangle_{XY}$ evolves to $|N, 0\rangle_{XY} + \exp(iN\phi)|0, N\rangle_{XY}$. This gives rise to interference oscillations N times as fast as those of a single photon with the same frequency (2).

In the experimental results (Fig. 3), two- to five-photon coincidence rates were measured as a function of the MZ phase. We denote the number of photon “clicks” in detectors D_1 and D_2 as N_1 and N_2 , respectively. The N_1, N_2 coincidence rate is expected to demonstrate a de Broglie wavelength (6) of λ/N , where $N = N_1 + N_2$ is the total number of photons. The red curves are produced by an analytical model of the experiment, accounting for the overall transmission and the positive operator-value measures of the detectors (26). The two- and three-photon results were measured simultaneously with $\gamma_2 = \gamma_3 = 1$ (the subscript of γ denotes the value of N for which it was optimized) using a mere 2 mW of ultraviolet power to pump the SPDC. The results in Fig. 3, C and D, show a visibility of $86 \pm 0.6\%$ for 2, 1 photons and $80 \pm 1.9\%$ for 3, 0 photons. The four-photon measurements were taken with $\gamma_4 = \sqrt{3}$ and 25 mW of ultraviolet pump power. They exhibit similar visibilities for the 3, 1 and the 2, 2 options: $74 \pm 3\%$ and $73 \pm 2.4\%$, respectively (Fig. 3, E and F). Finally, the $N = 5$ NOON state has a visibility of $V = 42 \pm 2\%$ for 3, 2 photons and was taken using 215 mW of ultraviolet pump power and setting $\gamma_5 = 9/(\sqrt{10} + 1) \approx 2.16$ (Fig. 3G), implying that $\gamma_5^2 \approx 4.7$ times as many photon pairs originate from the coherent state as from the SPDC. All the above visibilities manifest the high NOON-state overlap of the generated states. These visibilities significantly exceed the super-resolution bound for classical states, which we have recently derived (27). In particular, the classical bound for the five-photon measurement is 16.67%, which is surpassed here by more than 12 standard deviations. The visibility of the experimental plots is determined by the overall setup transmission, which we denote as η . For $\eta < 1$, high-order events contribute a background to the N -fold coincidence

rate. In the current setup we estimate $\eta \approx 0.12$ on the basis of the SPDC coincidence-to-singles ratio.

The realized scheme works naturally for arbitrary N (24). This requires no alterations to the setup, except for the use of detectors that can resolve N -photon events and setting γ appropriately. This is in contrast to previous experiments that were customized to a specific value of N (16–18). Our implementation is limited to $N = 5$ mainly by the overall setup transmission ($\eta \approx 0.12$). A higher value would allow the generation of even larger states (26). Furthermore, most of the photons in this scheme originate from the coherent (classical) light source, which is practically unlimited in intensity. This eliminates the need for bright SPDC sources, providing experimental simplification. Finally, the scheme involves no state projection or post-selection, implying that all the N -photon events contribute to the measurable N -photon interferences.

The N -fold coincidence plots of Fig. 1D exhibit N zero points, as expected from perfect maximally path-entangled states, albeit with somewhat modulated peak heights. In fact, it has been shown (24) that as a result of the high fidelity, we can expect a phase sensitivity that is only slightly lower than the Heisenberg limit for a given N -photon component. Thus, the small deviation from an ideal NOON state is not a major limiting factor and is more than compensated for by the intrinsic 100% efficiency. Furthermore, although the state at hand is optimized for a given $N = N_0$, there is a range of N values surrounding N_0 that also have considerable NOON fidelity (Fig. 1, A and C). Thus, the generated state is actually a superposition of NOON states with various photon numbers, each of which contributes to the enhanced phase sensitivity. As a result, the current method could allow exceeding of the standard quantum limit while accounting for the complete photon number distribution (25).

It is interesting that a setup similar to the one in Fig. 2A, but with much stronger light fields, is commonly used for obtaining quantum noise reduction via homodyne detection (6). Homodyne detection is a highly developed technique based on the measurement of the continuous-variable field quadratures. Our experiment shows that extending the concept of squeezed vacuum and coherent light interference into the weak local oscillator regime is extremely fruitful. The states emerging from this interference, for a given N , are almost perfect NOON states. This implies a fundamental connection between quantum noise reduction (with continuous variables) and creation of NOON states (in photon number-resolving experiments).

The key to extending this work to even higher NOON states is in improving the overall transmission. This highlights the need for high-purity SPDC sources that can be spectrally mode-

matched to coherent states (28–30). Improved single-mode coupling of the photon pairs and high-efficiency photon number-resolving detectors (31, 32) would also be extremely beneficial. The inherent scalability opens the way to exciting applications in high-sensitivity interferometry, quantum imaging, and sub-Rayleigh lithography.

References and Notes

1. M. Nielsen, I. Chuang, *Quantum Computation and Quantum Information* (Cambridge Univ. Press, Cambridge, 2000).
2. J. Dowling, *Contemp. Phys.* **49**, 125 (2008).
3. E. Schrödinger, *Naturwissenschaften* **23**, 807 (1935).
4. L. Aolita, R. Chaves, D. Cavalcanti, A. Acín, L. Davidovich, *Phys. Rev. Lett.* **100**, 080501 (2008).
5. H. S. Eisenberg, G. Khoury, G. A. Durkin, C. Simon, D. Bouwmeester, *Phys. Rev. Lett.* **93**, 193901 (2004).
6. C. C. Gerry, P. L. Knight, *Introductory Quantum Optics* (Cambridge Univ. Press, Cambridge, 2005).
7. A. N. Boto et al., *Phys. Rev. Lett.* **85**, 2733 (2000).
8. J. A. Jones et al., *Science* **324**, 1166 (2009).
9. Y.-A. Chen et al., *Phys. Rev. Lett.* **104**, 043601 (2010).
10. K. T. McCusker, P. G. Kwiat, *Phys. Rev. Lett.* **103**, 163602 (2009).
11. K. T. Kapale, J. P. Dowling, *Phys. Rev. Lett.* **99**, 053602 (2007).
12. H. Cable, J. P. Dowling, *Phys. Rev. Lett.* **99**, 163604 (2007).
13. G. J. Pryde, A. G. White, *Phys. Rev. A* **68**, 052315 (2003).
14. J. Fiurásek, *Phys. Rev. A* **65**, 053818 (2002).
15. H. F. Hofmann, *Phys. Rev. A* **70**, 023812 (2004).
16. M. D'Angelo, M. V. Chekhova, Y. Shih, *Phys. Rev. Lett.* **87**, 013602 (2001).
17. M. W. Mitchell, J. S. Lundeen, A. M. Steinberg, *Nature* **429**, 161 (2004).
18. H. Kim, H. S. Park, S.-K. Choi, *Opt. Express* **17**, 19720 (2009).
19. P. Walther et al., *Nature* **429**, 158 (2004).
20. B. H. Liu et al., *Phys. Rev. A* **77**, 023815 (2008).
21. K. J. Resch et al., *Phys. Rev. Lett.* **98**, 223601 (2007).
22. R. Okamoto et al., *N. J. Phys.* **10**, 073033 (2008).
23. T. Nagata, R. Okamoto, J. L. O'Brien, K. Sasaki, S. Takeuchi, *Science* **316**, 726 (2007).
24. H. F. Hofmann, T. Ono, *Phys. Rev. A* **76**, 031806 (2007).
25. L. Pezzé, A. Smerzi, *Phys. Rev. Lett.* **100**, 073601 (2008).
26. See supporting material on Science Online.
27. I. Afek, O. Ambar, Y. Silberberg, *Phys. Rev. Lett.* **104**, 123602 (2010).
28. P. J. Mosley et al., *Phys. Rev. Lett.* **100**, 133601 (2008).
29. P. J. Mosley, J. S. Lundeen, B. J. Smith, I. A. Walmsley, *N. J. Phys.* **10**, 093011 (2008).
30. A. Valencia, A. Ceré, X. Shi, G. Molina-Terriza, J. P. Torres, *Phys. Rev. Lett.* **99**, 243601 (2007).
31. C. Silberhorn, *Contemp. Phys.* **48**, 143 (2007).
32. A. E. Lita, A. J. Miller, S. W. Nam, *Opt. Express* **16**, 3032 (2008).
33. We thank B. Dayan for helpful discussions. Supported by an Ilan Ramon Fellowship (I.A.) and the German-Israeli Foundation for Scientific Research and Development.

Supporting Online Material

www.sciencemag.org/cgi/content/full/328/5980/879/DC1
SOM Text

Fig. S1

References

10 February 2010; accepted 13 April 2010
10.1126/science.1188172

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

The following resources related to this article are available online at www.sciencemag.org (this information is current as of May 13, 2010):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/328/5980/882>

Supporting Online Material can be found at:

<http://www.sciencemag.org/cgi/content/full/328/5980/882/DC1>

This article **cites 16 articles**, 3 of which can be accessed for free:

<http://www.sciencemag.org/cgi/content/full/328/5980/882#otherarticles>

This article appears in the following **subject collections**:

Chemistry

<http://www.sciencemag.org/cgi/collection/chemistry>

Hydrogen Bonding Controls the Dynamics of Catechol Adsorbed on a $\text{TiO}_2(110)$ Surface

Shao-Chun Li,^{1*} Li-Na Chu,² Xue-Qing Gong,² Ulrike Diebold^{1,3*}

Direct studies of how organic molecules diffuse on metal oxide surfaces can provide insights into catalysis and molecular assembly processes. We studied individual catechol molecules, $\text{C}_6\text{H}_4(\text{OH})_2$, on a rutile $\text{TiO}_2(110)$ surface with scanning tunneling microscopy. Surface hydroxyls enhanced the diffusivity of adsorbed catecholates. The capture and release of a proton caused individual molecules to switch between mobile and immobile states within a measurement period of minutes. Density functional theory calculations showed that the transfer of hydrogen from surface hydroxyls to the molecule and its interaction with surface hydroxyls substantially lowered the activation barrier for rotational motion across the surface. Hydrogen bonding can play an essential role in the initial stages of the dynamics of molecular assembly.

The diffusion of molecules across surfaces is key to a wide variety of processes, including catalysis, molecular self-assembly, and the formation of supramolecular structures (1, 2). Direct observation of single diffusion events has given many insights into atomic-scale processes, such as the occurrence of long jumps (3). One particular interesting phenomenon, which is best viewed by scanning tunneling microscopy (STM), is the interaction between neighboring molecules or coadsorbates (4–8). For example, metal oxide surfaces are usually hydroxylated or hydrated under ambient conditions, so hydrogen diffusion is particularly important. Its influence on the diffusion processes of other adsorbates has been established with area-averaging techniques (9), but atomic-scale information about how exactly it affects diffusion mechanisms is sparse.

We used STM to study the dynamics of catechol [$\text{C}_6\text{H}_4(\text{OH})_2$] adsorbed on rutile $\text{TiO}_2(110)$, which serves as a model for individual organic molecules adsorbed on an oxide surface. TiO_2 is a wide-band-gap (~ 3 eV) semiconductor that finds applications in a wide variety of technical fields, and the rutile $\text{TiO}_2(110)$ surface has evolved as a main model system in surface science for studying molecular processes on oxides (10–13). Catechol is an aromatic chemical compound with two neighboring hydroxyl groups substituted on the benzene ring (Fig. 1A). It is a common linker molecule for anchoring organics on surfaces; for example, dyes for dye-sensitized solar cells.

The $\text{TiO}_2(110)$ surface consists of rows of fivefold-coordinated Ti atoms (Ti_{5c}) separated by rows of twofold-coordinated, bridge-bonded O atoms (O_{2c}) (10). A few percent of these O_{2c} atoms can be removed by heating under vacuum. The

number of the resulting O vacancies is somewhat variable depending on sample pretreatment. Dissociation of water at these defects leads to a partially hydroxylated surface. Figure 1B shows an STM image with a small amount of O vacancies, hydroxyls, and a few catechol molecules. Each species is distinguished by its apparent height in STM (Fig. 1C).

When a catechol molecule adsorbs on $\text{TiO}_2(110)$, both terminal OH groups dissociate and the H's bind to the O_{2c} atoms to form additional surface hydroxyls. The molecule itself is linked to two neighboring Ti_{5c} sites via its O atoms; that is, as bridging bi-dentate catecholate (14). Such a bridging bi-dentate configuration is also common to a large variety of organic acids on $\text{TiO}_2(110)$ (10, 15). For low surface coverages, when the catecholate molecules are isolated from each other, the benzene ring is upright with its plane parallel to the Ti_{5c} rows. (The molecules are still imaged as round protrusions in STM.) At the saturation coverage of half a monolayer (ML, where 1 ML corresponds to the number density of Ti_{5c} sites), a closely packed layer with a (4×1) superstructure is formed, and neighboring molecules tilt to the left and right by $\sim 30^\circ$ with respect to the Ti_{5c} row in an alternating fashion (14). The H atom can easily transfer back to the catecholate from the

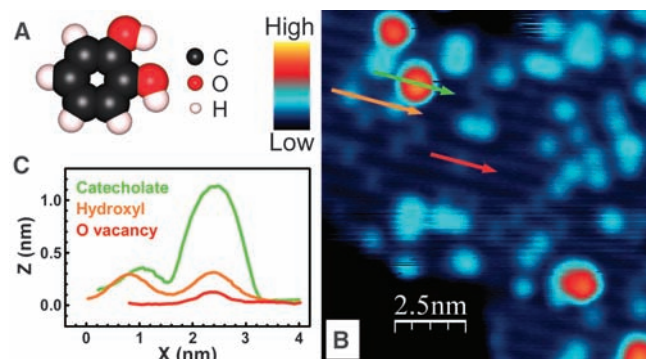
surface O_{2c} (activation energy E_{act} is calculated to be 0.16 eV only) (14). In fact, in the (4×1) overlayer, half of the catecholates have regained one H and form a H bond with a neighboring molecule. (14) The facile H transfer between adsorbed catecholates and the surface is instrumental for molecular diffusion across the $\text{TiO}_2(110)$ surface.

We followed the diffusion of isolated catecholate molecules (coverages below 0.05 ML) across the surface in time-lapse STM images. The $\text{TiO}_2(110)$ surface forms a strongly directional template; statistics on a few hundred different events showed preferred motion along Ti_{5c} rows. Cross-row jumps were seldom observed. All of the catecholate molecules exhibited similar STM morphologies but were present in two distinct states according to their diffusion rates. Some of the molecules stayed still during the measurement time, whereas others hopped between successive images (typically the scan speed was ~ 1 to 2 min per image), or changed their position even during one scan line, giving rise to molecules that appeared to be cut in half (Fig. 2, fig. S1, and movie S1) (16). We refer to these two different types of species as immobile and mobile catechol, respectively.

The diffusion rate of mobile catecholates is estimated to be at least one order of magnitude greater than that of immobile molecules. We tracked the number and position of hydroxyls in the vicinity of such mobile and immobile catecholate molecules and found that in the case of mobile catecholate, extra H played the key role. For example, in Fig. 2B, a catechol molecule switches from one bi-dentate position (marked with the two black spots in Fig. 2A) to another one, two lattice sites to the right, revealing a hidden H atom (red dot) that was not distinguishable from the catechol in Fig. 2A. The molecule stayed in its original position for ~ 1 min (Fig. 2C) and then hopped to a new position (Fig. 2D) one lattice site to the left. It also left behind an additional H (red dot); no further hopping was observed in consecutive images.

Another such example is shown in fig. S2 (see also movie S2) (16), where two H atoms in the vicinity of the catecholate molecule appear to be involved in the hopping. After the molecule's

Fig. 1. (A) Ball model of a catechol molecule. (B) STM image (10 by 12 nm, sample bias voltage = +1.48 V, tunneling current = 0.05 nA, temperature = 300 K) of a $\text{TiO}_2(110)$ surface with a low coverage of catechol. Lines point out three species with different heights; the respective profiles are displayed in (C). (C) The highest features (green line) are due to single catechol molecules; the second-highest (orange line) are due to surface hydroxyls, some of which originate from the dissociation of water on bridge-bonded O vacancies on $\text{TiO}_2(110)$, which appear as the lowest protrusions (red line). Z, apparent height; X, distance.



¹Department of Physics and Engineering Physics, Tulane University, New Orleans, LA 70118, USA. ²Labs for Advanced Materials, Research Institute of Industrial Catalysis, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, People's Republic of China. ³Institute of Applied Physics, Vienna University of Technology, Wiedner Hauptstrasse 8–10, Vienna, Austria.

*To whom correspondence should be addressed. E-mail: sli1@tulane.edu (S.L.); diebold@iap.tuwien.ac.at (U.D.)

hop to the right is completed, the H atom to the left has switched its position from a lower O_{2c} row to the higher one (green arrow in fig. S2A), indicating a counterclockwise rotation of the catecholate (red arrow in fig. S2A) (16). The second H moves together with the catecholate in fig. S2, B and C, while being split to the left in fig. S2D (16).

Such H motion was never observed in the vicinity of an immobile catecholate, as seen in movies S1 and S2 (16).

The STM results reveal that H is directly involved in the motion of catechol molecules, indicating that the mobility and position of the H on the $TiO_2(110)$ surface must affect the diffusion

Fig. 2. Snapshots of an STM movie of a mobile catecholate molecule on $TiO_2(110)$. The time interval between each frame is ~ 1.0 min, and the data were taken at room temperature. The upper and lower panels show the same images with and without an overlay of the substrate lattice grid, respectively; the orange feature at the lower left corner is the next terrace of the $TiO_2(110)$ substrate. (A) The two black dots in the overlay represent the Ti_{5c} lattice sites occupied by bi-dentate catecholate; the yellow dots indicate hydroxyl groups. (B) The catecholate moves to the right during a few scan lines, revealing an additional hydroxyl (red dot). (C) The catecholate returns to its original position. (D) The catecholate has moved one lattice site to the left. It left behind an additional H (red dot) and stayed immobile in subsequent STM images.

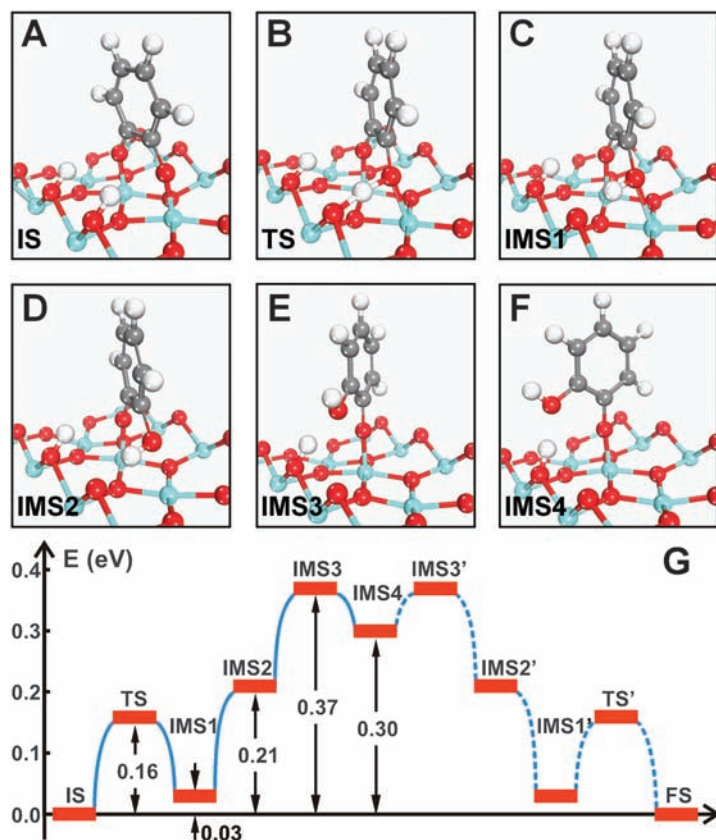


Fig. 3. Calculated rotational (clockwise) pathway of a catechol on $TiO_2(110)$ with the lowest E_{act} . (A) Initial state of a catecholate adsorbed on a $TiO_2(110)$ surface: The catechol is fully dissociated, and the H's split off from both the OH groups sit on the neighboring O row. (B) TS: One $O_{catechol}-Ti_{5c}$ bond is broken and a H atom is back-transferred from one of the O_{2c} atoms. (C to F) The following IMSs during the first half of the rotation. (G) Energy profile of the rotational pathway with the corresponding key states [(A) to (F)]. The second half of the rotation (dashed line) is just the time reversal of the first one. FS, final state.

of the catecholates. The intrinsic diffusion of hydroxyls along O_{2c} rows is rather low at room temperature [hopping rate $\sim 1 \times 10^{-5} s^{-1}$ (17)], whereas it is known that weakly adsorbed molecular water can drive H cross-row diffusion on the surface (13). The water background pressure was carefully suppressed during our experiments, so that the water-induced cross-row diffusion was prohibited. (The rest of the hydroxyls in Fig. 2 and fig. S2 were all immobile during the measurement.) However, we could purposely introduce additional water into the chamber (partial pressure $< 1 \times 10^{-11}$ mbar) to change the distribution of surface hydroxyls. Figure S3 shows the same area before (fig. S3A) and after (fig. S3B) additional water was dosed onto the sample surface (16). Many of the hydroxyls switched their positions because of water-induced cross-row diffusion (indicated by the arrows in fig. S3B) (16). Again, one additional hydroxyl became visible, because the water had driven it away from the catecholate molecule (red dot). This catecholate switched from a mobile to an immobile state after releasing the hidden H.

To obtain a detailed understanding of the catechol diffusion on $TiO_2(110)$, we simulated this process by performing density functional theory calculations. In particular, a constrained minimization scheme was used to determine the transition states (TSs) and estimate corresponding E_{act} 's. The adsorption structure with the highest binding energy (1.15 eV with respect to a free catechol molecule) is a fully dissociated, bridging bi-dentate catecholate (14), which has donated its two H atoms to nearby O_{2c} atoms (Fig. 3A). Starting with this initial state (IS), different diffusion routes were considered for the catecholate to move one lattice site away from the original position. We found that a rotational route is energetically the most feasible, where the molecule rotates around an $O_{catechol}-Ti_{5c}$ bond. Other paths, such as a translation where the catecholate assumes a bi-dentate chelating intermediate state, with both catechol O atoms bound to the same surface Ti_{5c} atom, have unreasonably high E_{act} 's, which lie even above the adsorption energy.

The first half of the pathway with the lowest E_{act} is depicted in Fig. 3. In the first step, one $O_{catechol}-Ti_{5c}$ bond is broken and a H atom is back-transferred from one of the O_{2c} atoms. This leads to a TS (Fig. 3B) that is 0.16 eV higher in energy than the bi-dentate bridging catecholate in the IS. In contrast, just breaking the $O_{catechol}-Ti_{5c}$ bond alone would lead to a TS with a much higher energy (1.22 eV with respect to the IS); thus, the $O_{catechol}$ needs to be terminated with a H in order to "lift one leg" from the surface. In subsequent steps, the catechol rotates around the remaining $O_{catechol}-Ti_{5c}$ bond. As we show in Fig. 3, this rotation process consists of several intermediate states (IMSs) (Fig. 3, C to F) without a clear TS, and the energy of the IMS with the highest energy, IMS3 (Fig. 3E), is only 0.37 eV less stable than the IS. Such an activation barrier can be easily overcome at room temperature.

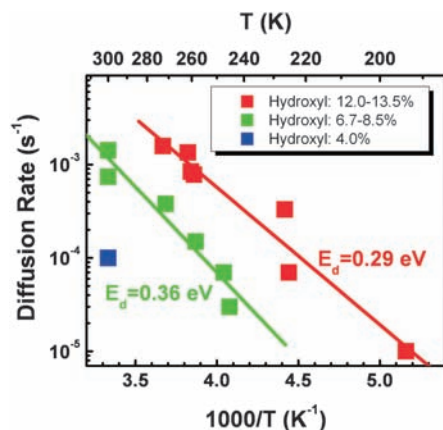


Fig. 4. Arrhenius plots of the diffusion of catechol on $\text{TiO}_2(110)$ surfaces with different hydroxyl coverages in the temperature (T) range between 195 and 300 K. Squares in different colors represent the diffusion rate for surfaces with different hydroxyl coverages. Assuming a thermally activated process for the hopping rate, $h = \text{vexp}(-E_d/k_B T)$, the data in red and green give activation barriers E_d of 0.29 ± 0.03 and 0.36 ± 0.05 eV, respectively.

The remaining H on the O_{2c} surface atom is also critical for this motion. In the clockwise rotation in Fig. 3, the catechol interacts with the remaining surface hydroxyl during IMS3 and IMS4 (Fig. 3E, F), forming an intermediate $\text{H}_{\text{hydroxyl}}\text{-OH}_{\text{catechol}}$ entity. In a scenario where the molecule rotates counterclockwise and faces the nonhydroxylated O_{2c} row in Fig. 3, the overall estimated activation barrier for the IMS is much higher (0.72 eV, fig. S4). During such a rotation, the catechol even loses its additional H atom and forms a bi-dentate chelating configuration at a Ti_{5c} atom as an IMS (IMS5, fig. S4G). The formation of the $\text{H}_{\text{hydroxyl}}\text{-OH}_{\text{catechol}}$ entity facilitates the rotation of catechol. These calculations support the experimental evidence that the hydroxylation of surface O atoms is instrumental for the catechol diffusion across the surface.

The temperature dependence of the diffusion rate was measured for the surfaces with various hydroxyl coverages. The Arrhenius plot in Fig. 4 shows that the diffusion rate is weakly dependent on the temperature, but a few percent increase in the hydroxyl concentration drastically enhanced the diffusion rate. Fitting these data and assuming a thermally activated (over-the-barrier) process, $h = \text{vexp}(-E_d/k_B T)$ [where h is the hopping (diffusion) rate, v is the pre-factor, k_B is the Boltzmann constant, and T is the temperature] gives an effective diffusion barrier E_d between ~ 0.36 and ~ 0.29 eV for two different hydroxyl coverages (6.7 to 8.5% and 12.0 to 13.5%). This value is in good agreement with the theoretical value of 0.37 eV. The pre-factor v is very low ($\sim 10^3 \text{ s}^{-1}$) but is similar to values shown in recent reports on other complex systems (6–8).

Our results show that the degree of hydroxylation of a surface is key for the facile diffusion of catechol across the surface. The effect is

different from that of pyridine on high-surface-area MgO powder (9), for example, where water displaces molecules from stronger to weaker adsorption sites. In our case, the molecule stays at the same position but switches mobility by gaining or losing an extra H atom, which may enable the molecule to roll over to the next lattice site. Moreover, our calculations indicate that only rotational motion is feasible at room temperature and that the catechol has to release the H atom and regain it in order to progress across the surface; otherwise, only a back-and-forth motion between two neighboring lattice sites would be possible. This situation is somewhat similar to the roll-over motion of water dimers that was observed on $\text{Pt}(111)$ (4) and $\text{TiO}_2(110)$ (7). To the best of our knowledge, however, this is the first atomic-scale study that shows the importance of H for the diffusion of organics across a surface.

References and Notes

1. T. Trevethan, A. L. Shluger, *J. Phys. Chem. C* **112**, 19577 (2008).
2. J. V. Barth, *Surf. Sci. Rep.* **40**, 75 (2000).
3. M. Schunack et al., *Phys. Rev. Lett.* **88**, 156102 (2002).
4. T. Mitsui, M. K. Rose, E. Fomin, D. F. Ogletree, M. Salmeron, *Science* **297**, 1850 (2002).

5. S. Horch et al., *Nature* **398**, 134 (1999).
6. B. G. Briner, M. Doering, H. P. Rust, A. M. Bradshaw, *Science* **278**, 257 (1997).
7. J. Matthiesen et al., *Phys. Rev. Lett.* **102**, 226101 (2009).
8. K. L. Wong et al., *Science* **315**, 1391 (2007).
9. X. Wang et al., *J. Phys. Chem. C* **113**, 2228 (2009).
10. U. Diebold, *Surf. Sci. Rep.* **48**, 53 (2003).
11. C. Lun Pang, R. Lindsay, G. Thornton, *Chem. Soc. Rev.* **37**, 2328 (2008).
12. Z. Zhang, O. Bondarchuk, B. D. Kay, J. M. White, Z. Dohnalek, *J. Phys. Chem. B* **110**, 21840 (2006).
13. S. Wendt et al., *Phys. Rev. Lett.* **96**, 066107 (2006).
14. S.-C. Li et al., *J. Am. Chem. Soc.* **131**, 980 (2009).
15. H. Onishi, Y. Iwasawa, *Chem. Phys. Lett.* **226**, 111 (1994).
16. See supporting material on Science Online.
17. S.-C. Li et al., *J. Am. Chem. Soc.* **130**, 9080 (2008).
18. S.L. and U.D. acknowledge financial support by the U.S. Department of Energy (grant DE-FG02-05ER15702) and by NSF (grant CHE-0715576). X.Q.G. is grateful for financial support from the National Nature Science Foundation of China (grant 20703017) and the 111 project (grant B08021). Valuable discussions with A. Selloni are gratefully acknowledged.

Supporting Online Material

www.sciencemag.org/cgi/content/full/328/5980/882/DC1

Materials and Methods

Figs. S1 to S4

Movies S1 to S3

References

16 February 2010; accepted 15 April 2010

10.1126/science.1188328

Heterogeneous Accretion and the Moderately Volatile Element Budget of Earth

M. Schönbächler,^{1,2*} R. W. Carlson,² M. F. Horan,² T. D. Mock,² E. H. Hauri²

Several models exist to describe the growth and evolution of Earth; however, variables such as the type of precursor materials, extent of mixing, and material loss during accretion are poorly constrained. High-precision palladium-silver isotope data show that Earth's mantle is similar in $^{107}\text{Ag}/^{109}\text{Ag}$ to primitive, volatile-rich chondrites, suggesting that Earth accreted a considerable amount of material with high contents of moderately volatile elements. Contradictory evidence from terrestrial chromium and strontium isotope data are reconciled by heterogeneous accretion, which includes a transition from dominantly volatile-depleted to volatile-rich materials with possibly high water contents. The Moon-forming giant impact probably involved the collision with a Mars-like protoplanet that had an oxidized mantle, enriched in moderately volatile elements.

Planetary accretion, core formation, and the loss of volatile elements shaped the compositional evolution of Earth. Radioactive decay systems (for instance, Hf-W and U-Pb) have been widely used to constrain the timing of these early processes (1–9). Resolving the composition and volatile content of the material from which Earth accumulated provides constraints on the conditions that lead to Earth's formation. Here, we report high-precision Ag isotope data for unaltered mantle-derived igneous rocks from

a variety of geological settings (10) to represent the bulk silicate Earth (BSE). These data provide the basis for accretion models that combine the results from the short-lived chronometer ^{107}Pd - ^{107}Ag with those of the ^{53}Mn - ^{53}Cr , ^{87}Rb - ^{87}Sr , and ^{182}Hf - ^{182}W decay systems. Except for Rb-Sr, terrestrial core formation influences all of these decay systems, because parent and daughter elements partition differently into metal and silicates. The Pd-Ag, Mn-Cr, and Rb-Sr chronometers also track volatility-related processes, because Ag, Rb, and Mn possess lower condensation temperatures than Pd, Sr, and Cr (11, 12). The combined application of these decay systems provides constraints on the nature and timing of volatile-element depletion experienced during the accretion of Earth or by Earth's precursor materials. For

¹School of Earth, Atmospheric and Environmental Sciences, University of Manchester, Manchester M13 9PL, UK. ²Department of Terrestrial Magnetism, Carnegie Institution of Washington, 5241 Broad Branch Road, NW, Washington, DC 20015, USA.

*To whom correspondence should be addressed: E-mail: maria.schonbachler@manchester.ac.uk

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

The following resources related to this article are available online at www.sciencemag.org (this information is current as of May 13, 2010):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:
<http://www.sciencemag.org/cgi/content/full/328/5980/884>

Supporting Online Material can be found at:
<http://www.sciencemag.org/cgi/content/full/328/5980/884/DC1>

This article **cites 30 articles**, 1 of which can be accessed for free:
<http://www.sciencemag.org/cgi/content/full/328/5980/884#otherarticles>

This article appears in the following **subject collections**:
Geochemistry, Geophysics
http://www.sciencemag.org/cgi/collection/geochem_phys

example, a successful way to obtain consistent Hf-W and Pd-Ag accretion ages is to vary the volatile content of the terrestrial building blocks. Heterogeneous accretion, where the composition of the material from which Earth was built varies with time, is a model that was proposed more than 30 years ago (13) and provides a means to explain seemingly contradictory radiometric data regarding early Earth materials.

All accretion models constrained by the Pd-Ag decay system require the knowledge of specific parameters such as the initial abundance of ^{107}Pd in the solar system, the Ag isotope composition, and the Pd/Ag ratio of BSE and the precursor materials from which Earth accreted (14). Data from iron meteorites (11) and carbonaceous chondrites indicate a solar system initial $^{107}\text{Pd}/^{108}\text{Pd}$ ratio in the range of $5.9 (\pm 2.2) \times 10^{-5}$ (15). Based on our mantle-derived samples, we estimate a BSE Ag isotope composition of $\epsilon^{107}\text{Ag} = -2.2 \pm 0.7$ (2 SD), relative to the National Institute of Standards and Technology SRM978a Ag standard (table S1). This value overlaps with that of primitive carbonaceous chondrites (15).

Cadmium serves as a proxy to evaluate whether the similarity in Ag isotope composition between BSE and carbonaceous chondrites is a fortuitous result of mass-dependent fractionation. Cadmium and Ag isotopes have similar masses and are both mainly chalcophile. Silver, however, has a higher half-mass condensation temperature [996 K (12)] than Cd [652 K (12)], which implies that mass-dependent isotope fractionation due to volatility affects Ag less than Cd. This concept is supported by data from ordinary chondrites, which show moderate fractionations in Ag isotope composition accompanied by substantial Cd isotope variations that are one order of magnitude larger (14, 15). Because the Cd isotope compositions of the carbonaceous chondrites Orgueil, Murchison, and Allende are identical to that of the BSE within analytical uncertainty (14, 16), there is no basis to expect appreciable Ag isotope fractionation between carbonaceous chondrites and Earth. Therefore, we use the C-chondrite-terrestrial mantle similarity to model the accretion history of Earth.

Most of the BSE elemental abundances suggest that the materials that built up Earth were similar to carbonaceous chondrites (17–19). Although moderately volatile elements are more depleted in Earth than in any known carbonaceous chondrite, Earth's composition falls on the extension of the volatile-depletion trend defined by carbonaceous chondrites (17, 19). This depletion trend follows the condensation sequences calculated for the solar nebula (12, 20) and is probably established early in the solar nebula. Moreover, chronometers like Mn-Cr (19) and Rb-Sr (21, 22) indicate that Earth accreted from planetesimals with considerable depletions in volatile elements. For the purposes of modeling terrestrial accretion and core formation, we assume, as a first approximation, that Earth's start-

ing materials had an average volatile content similar to CV3 chondrites, the most volatile-depleted carbonaceous chondrite group [not to be mistaken with direct accretion of CV3 chondrites (10)]. A simple model based on Pd-Ag systematics, in which terrestrial core formation is instantaneous and segregation of Pd to the core is complete [two-stage model (10)], suggests that the final metal-silicate equilibration occurred less than ~9 million years (My) after the formation of the solar system (Fig. 1A). More realistic models, in which accretion and core formation are continuous and the newly accreted material equilibrates only with BSE, yield similar short durations for accretion and core segregation (Fig. 1B). Greater extents of volatile depletion in the source materials for Earth lead to even shorter accretion ages and imply that Earth inherited its volatile-element depletion almost exclusively from its source materials.

The short Pd-Ag accretion ages contrast with those of the Hf-W decay system, where the same models yield an age of ~30 My (7–9). To understand the discrepant Pd-Ag and Hf-W model ages, we consider the possibility of incomplete metal-silicate equilibration between the BSE and the newly accreted material. Iron meteorite studies indicate that planetesimals segregated into a silicate mantle and metallic core within the first few million years of solar system history (11, 15, 23), thereby suggesting that Earth accreted from already-differentiated planetesimals. The degree to which the metallic cores of such objects would equilibrate with Earth's mantle on impact is still debated (1, 5). Metal-silicate disequilibrium during accretion and core formation has a strong effect on model ages obtained from the Hf-W decay system; large extents of

disequilibrium throughout accretion can yield model ages greater than 100 My for the final accretion stages (2, 6). If Earth accreted from volatile-depleted (high Pd/Ag ratio) material, then no combination of accretion age and degree of metal-silicate equilibration that suits the Hf-W data can simultaneously explain the Pd-Ag constraints (Fig. 2A). Metal-silicate disequilibrium has a muted effect on the growth of ^{107}Ag (in comparison to the ^{182}Hf - ^{182}W system) because of the lower initial abundance of ^{107}Pd in the solar system, whereas both Pd-Ag and Hf-W fractionation are of similar magnitude during metal-silicate separation.

In contrast to Pd and Ag, Hf and W are both refractory elements that do not substantially fractionate with volatility (although W partitioning changes with oxidation state). Therefore, the Hf-W ages are much less sensitive to variations in the volatile content of the precursor material or to volatilization during the impact-driven accretion of Earth. The use of volatile-rich material CI-like chondrites as a terrestrial precursor for Pd-Ag models relaxes the time constraints on accretion and core formation (Fig. 3A). The overlapping Ag isotope compositions of the BSE and CI chondrites, together with the subchondritic Pd/Ag ratio of the BSE (due to the strong partitioning of the highly siderophile element Pd into the core), excludes only extremely rapid core formation during the first 3 My of solar system history. Although the accretion of volatile-rich material eliminates the apparent conflict between Hf-W and Pd-Ag systematics, it requires that volatile elements were lost at a later stage (for instance, during the Moon-forming giant impact) because the present-day Earth is depleted in those elements.

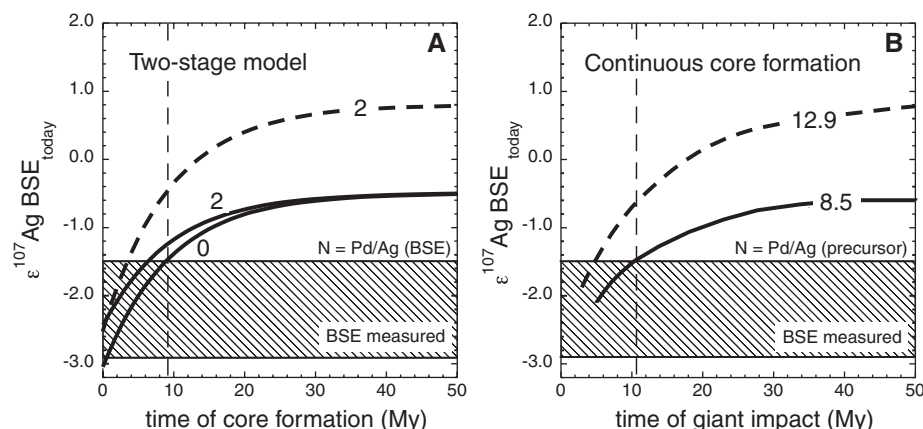


Fig. 1. The $\epsilon^{107}\text{Ag}$ of the present-day BSE is shown as a function of (A) the time of core formation for the two-stage model and (B) the time of the giant impact for the continuous core-formation model (10). In (A), the dashed curve is calculated for precursor materials with volatile-element depletion similar to Earth today (Pd/Ag = 12.9), whereas the two solid curves represent CV3-like precursor material (Pd/Ag = 8.5) and BSE Pd/Ag ratios of 0 and 2, respectively. The hatched region shows the measured Ag isotope composition of BSE (-2.2 ± 0.7), and the dashed vertical line indicates a core-formation age of ~9 My after the start of the solar system based on volatile-depleted precursor materials (CV-like) and a BSE Pd/Ag ratio of 0. For the continuous core-formation model (B), the vertical dashed line indicates that Earth has virtually completed its accretion and core formation within 11 My (time of the giant impact) if it was built from a volatile-depleted precursor (CV3-like, solid curve), and even earlier if the precursor material was as volatile-depleted as BSE today (dashed curve).

Several lines of evidence, however, indicate that, on average, Earth did not accrete solely from such volatile-rich material. For example, volatility influences the Mn-Cr decay system, as well as the Pd-Ag system. Applying core-formation models to the Mn-Cr system using CI starting material cannot match the ^{53}Cr composition of the BSE (Fig. 3A). The BSE possesses an $\epsilon^{53}\text{Cr}$ of about -0.4 (24, 25) or -0.2 (26, 27), relative to

most types of chondrites. The terrestrial ^{53}Cr deficit indicates that Earth accreted from material that became volatile-depleted very early in the lifetime of ^{53}Mn , possibly in the solar nebula (19). In addition, the precursor materials of the Moon, and presumably Earth, must have been volatile-depleted (low Rb/Sr ratio) within 20 My of solar system formation based on the low $^{87}\text{Sr}/^{86}\text{Sr}$ ratio found in old lunar crustal rocks (21, 22).

Thus, both the Mn-Cr and Rb-Sr chronometers demand volatile-depleted precursor materials relative to CI chondrites, whereas the Pd-Ag system requires a volatile-rich precursor to extend the Pd-Ag accretion ages to those suggested by Hf-W systematics.

Heterogeneous accretion may provide a solution to this paradox. According to our model calculations (Fig. 3), the accretion of volatile-depleted material (similar to today's volatile content), followed by $\sim 13\%$ of volatile-rich material toward the end of accretion, can reconcile the Mn-Cr, Rb-Sr, Pd-Ag, and Hf-W data for the terrestrial mantle. The exact percentage depends on the volatile content of the accreted material. Some volatile loss may have occurred during terrestrial accretion, and if so, the actual Mn/Cr or Ag/Pd ratios of the precursor material were higher (that is, more volatile-rich) than the ratios assumed in the model calculations (Fig. 3). A late veneer (28) that adds the final 1% of material after core formation ceased was included into the model (10) but does not notably affect the $\epsilon^{107}\text{Ag}$ of the BSE (^{107}Pd was extinct by the time the late veneer was added, and the amount of Ag in the late veneer is small compared with the BSE Ag content). Silver originating from the late veneer could dominate Earth's mantle if the late veneer delivered a much larger fraction than 1% of Earth's present mass; however, this is inconsistent with the abundances of highly siderophile elements in Earth's mantle (28) and would lead to Pd concentrations that are more than one order of magnitude higher than those observed today.

Dynamic modeling (29) indicates that a Moon-forming giant impact occurred toward the end of terrestrial accretion as the product of the collision between proto-Earth and a Mars-sized body (called Theia). Our model implies the late accretion of $\sim 13\%$ volatile-rich material, and this suggests that Theia may have resembled Mars, not only in size, but also in chemical composition [volatile-rich composition close to CI chondrites for Mn and Ag (10)], which is consistent with previous work (2). We incorporated a volatile-rich, Moon-forming giant impact into our model as the major source of the late-accreted volatile-rich material (Fig. 3). Within the uncertainty of the modeling, we cannot constrain whether the entire volatile-rich component accreted in a single event, such as the Theia collision, or over a more prolonged interval through several impacts.

The time sequence in which the different materials accreted is important. Early addition of some volatile-rich material is possible, but the exclusive addition of such material in the early accretion stage would result in an overabundance of radiogenic ^{107}Ag in the terrestrial mantle. In contrast, late accreted volatile-rich material contains less radiogenic ^{107}Ag due to the extended time period over which it developed with a low Pd/Ag ratio. Hence, our model explains the records of Ag, Cr, Sr, and W isotope composition of the BSE, if the accretion of volatile-

Fig. 2. (A) Applying the continuous core-formation model to the Hf-W chronometer, the time of the giant impact was determined as a function of the degree of metal-silicate equilibration (f). Curves are shown for $f = 0$ (no equilibration) to $f = 1$ (full equilibration) during the giant impact only (diamonds) and for disequilibrium ($f = 0.4$ to 1) during the entire accretion process (0 to 99%; crosses). **(B)** Each combination of Hf-W core-formation age (i.e., the time of the giant impact) and metal-silicate equilibration (f) obtained in (A) is used here as input parameters for the continuous core-formation model based on Pd-Ag [with $\text{Pd}/\text{Ag}_{(\text{precursor})} = 8.5$]. The Ag isotope composition of the BSE measured today (hatched area) does not overlap with any of the model results. For more details, see (10).

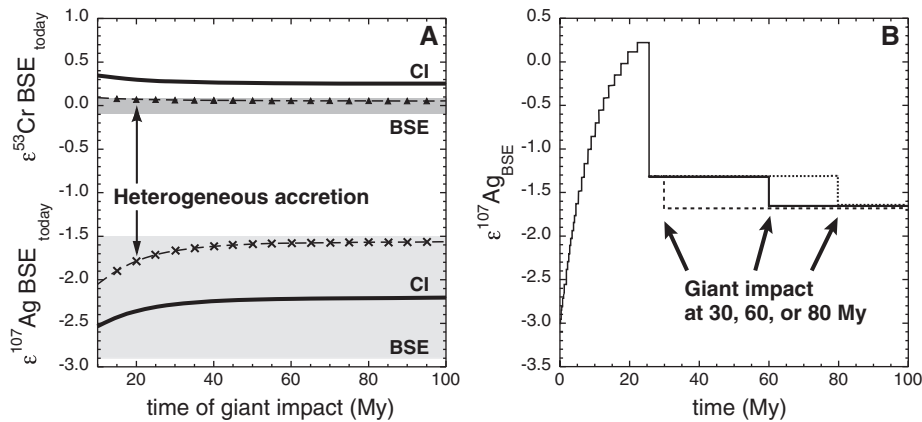
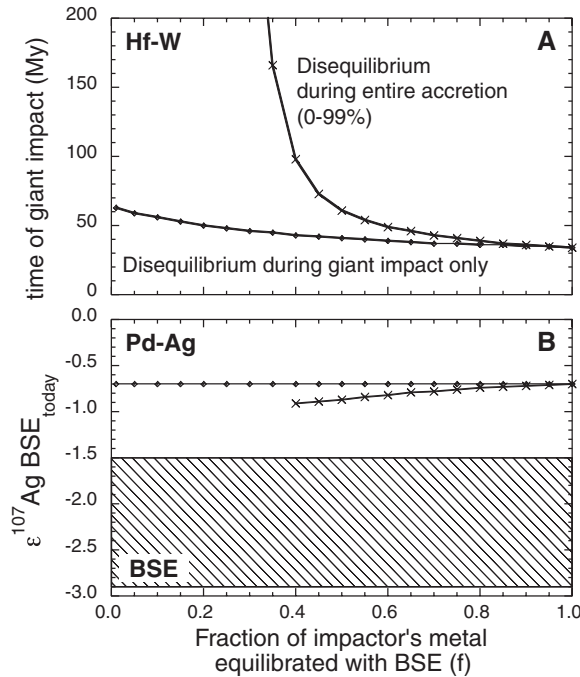


Fig. 3. The $\epsilon^{107}\text{Ag}$ and $\epsilon^{53}\text{Cr}$ of the BSE depend on the volatile depletion of the precursor material but vary only slightly with the accretion time scale. (A) A continuous core-formation model based on CI composition of the precursor materials (top solid curve) describing the $\epsilon^{53}\text{Cr}$ of BSE does not intersect the field defined by the measured Cr isotope composition of BSE (small gray area) and, therefore, is not a realistic solution. Dashed lines describe the results for heterogeneous accretion, which encompass the accretion of volatile-depleted material until Earth reaches 86% of its present-day size, followed by the addition of 4% of CI-like material and the giant impact, which added another 9% of volatile-rich, Mars-like material and a 1% late veneer with a Pd/Ag ratio of 4.1. The model curves for BSE Ag (crosses) and Cr (triangles) isotope composition overlap with the measured values of BSE (gray regions). **(B)** The Ag isotope evolution of the BSE over time for heterogeneous accretion, which includes the accumulation of 4% volatile-rich material when Earth reaches 86% of its current mass at ~ 26 My, leading to the first major drop in $\epsilon^{107}\text{Ag}$. The volatile-rich giant impact at 30 My reduces $\epsilon^{107}\text{Ag}$ further. Because all ^{107}Pd essentially decayed away by 30 My, there is little difference in the evolution of BSE Ag isotope composition if the giant impact happened at 30 My or if there was a time gap with no accretion followed by a giant impact at 60 or 80 My.

depleted material dominated the early accretion history of Earth.

In our heterogeneous accretion scenario (Fig. 3), 90% of Earth is assumed to have accreted within the first 30 My (3–5), and the Moon-forming giant impact (addition of 9% of Earth's mass) occurred at 60 My after the start of the solar system. The latter builds on the observation that Earth's mantle and the Moon possess identical W isotope compositions, whereas their Hf/W ratios are different (30). However, the timing of the giant impact (e.g., at 30, 60, or 80 My after the start of the solar system) does not change the model results for Pd-Ag (Fig. 3B) or Mn-Cr. Likewise, metal-silicate disequilibrium during accretion has no major consequences for the Pd-Ag and Mn-Cr model ages, because these chronometers respond less to the disequilibrium than Hf-W due to their lower initial abundances and the limited extent of Mn-Cr fractionation between metal and silicate.

An earlier heterogeneous accretion model (13), in which the early accretion of a reduced volatile-free component preceded a highly oxidized, volatile-rich component, links oxidized and volatile-rich material or, alternatively, reduced and volatile-poor material. This compositional link also holds up for Mars and some, but not all, meteorite classes. Based on the link between volatile content and oxidation state, our model is consistent with studies on elemental partitioning behavior (31, 32). These studies propose that the oxidation state of Earth's mantle changed from reduced to more oxidized toward the end of accretion to satisfy metal-silicate partition coefficients, the depletion in certain element concentrations (for example, Nb, V, and Cr), and the oxidation state of the present-day BSE (31, 32). However, we do not exclude the possibility that self-oxidation of the mantle also occurred and contributed to the changes in the oxygen fugacity of the mantle (32).

Recent dynamic models of accretion and planet formation also conclude that volatile-rich material accumulated toward the end of the terrestrial accretion (33, 34). For example, Earth acquired 1.6 to 38% of its volatile-rich material from outside 2.5 astronomical units toward the end of accretion (if Jupiter and Saturn started with nearly circular orbits) (34). This suggests that the material from which Earth accreted contained enough water so that we need no other sources of water (such as the late veneer) to account for Earth's oceans (33, 34). Still, we cannot entirely exclude an additional later delivery of water. The extent of water loss during impact-driven accretion and the amount of water stored in the mantle are uncertain, but <1% of CI material can supply Earth's total water content, as it is currently estimated (13, 35).

References and Notes

- C. J. Allègre, G. Manhès, C. Göpel, *Earth Planet. Sci. Lett.* **267**, 386 (2008).
- A. N. Halliday, *Nature* **427**, 505 (2004).
- A. N. Halliday, B. J. Wood, *Science* **325**, 44 (2009).
- S. B. Jacobsen, *Annu. Rev. Earth Planet. Sci.* **33**, 531 (2005).
- S. B. Jacobsen et al., *Philos. Trans. R. Soc. London Ser. A* **366**, 4129 (2008).
- T. Kleine, K. Mezger, H. Palme, C. Münker, *Earth Planet. Sci. Lett.* **228**, 109 (2004).
- T. Kleine, C. Münker, K. Mezger, H. Palme, *Nature* **418**, 952 (2002).
- R. Schoenberg, B. S. Kamber, K. D. Collerson, O. Eugster, *Geochim. Cosmochim. Acta* **66**, 3151 (2002).
- Q.-Z. Yin et al., *Nature* **418**, 949 (2002).
- Supporting material is available on Science Online.
- J. H. Chen, G. J. Wasserburg, in *Earth Processes: Reading the Isotopic Code*, Geophysical Monograph 95, A. Basu, S. R. Hart, Eds. (American Geophysical Union, Washington, DC, 1996), pp. 1–20.
- K. Lodders, *Astrophys. J.* **591**, 1220 (2003).
- H. Wänke, G. Dreibus, E. Jagoutz, in *Archaean Geochemistry*, A. Kröner, G. N. Hanson, A. M. Goodwin, Eds. (Springer, Berlin, 1984), pp. 1–24.
- F. Wombacher, M. Rehkämper, K. Mezger, C. Münker, *Geochim. Cosmochim. Acta* **67**, 4639 (2003).
- M. Schönbachler, R. W. Carlson, M. F. Horan, T. D. Mock, E. H. Hauri, *Geochim. Cosmochim. Acta* **72**, 5330 (2008).
- A. D. Schmitt, S. J. G. Galer, W. Abouchami, *Earth Planet. Sci. Lett.* **277**, 262 (2009).
- C. J. Allègre, G. Manhès, E. Lewin, *Earth Planet. Sci. Lett.* **185**, 49 (2001).
- W. F. McDonough, in *Treatise on Geochemistry*, Vol. 2: *The Mantle and Core*, R. W. Carlson, Ed. (Elsevier, Amsterdam, 2003), pp. 547–568.
- H. Palme, H. S. C. O'Neill, in *Treatise on Geochemistry*, Vol. 2: *The Mantle and Core*, R. W. Carlson, Ed. (Elsevier, Amsterdam, 2003), pp. 1–38.
- L. Schaefer, B. Fegley Jr., *Icarus* **205**, 483 (2010).
- R. W. Carlson, G. W. Lugmair, *Earth Planet. Sci. Lett.* **90**, 119 (1988).
- A. N. Halliday, D. Porcelli, *Earth Planet. Sci. Lett.* **192**, 545 (2001).
- A. Scherstén, T. Elliott, C. Hawkesworth, S. Russell, J. Masarik, *Earth Planet. Sci. Lett.* **241**, 530 (2006).
- F. Moynier, Q.-Z. Yin, B. Jacobsen, *Astrophys. J.* **671**, L181 (2007).
- A. Shukolyukov, G. W. Lugmair, *Earth Planet. Sci. Lett.* **250**, 200 (2006).
- L. Qin, C. M. O. D. Alexander, R. W. Carlson, M. F. Horan, T. Yokoyama, *Geochim. Cosmochim. Acta* **74**, 1122 (2010).
- A. Tringui, J. L. Birc, C. J. Allègre, C. Göpel, D. Ulfbeck, *Geochim. Cosmochim. Acta* **72**, 5146 (2008).
- C.-L. Chou, *Proc. Lunar Planet. Sci. Conf.* **9**, 219 (1978).
- R. M. Canup, E. Asphaug, *Nature* **412**, 708 (2001).
- M. Touboul, T. Kleine, B. Bourdon, H. Palme, R. Wieler, *Nature* **450**, 1206 (2007).
- H. St. C. O'Neill, *Geochim. Cosmochim. Acta* **55**, 1159 (1991).
- B. J. Wood, J. Wade, M. R. Kilburn, *Geochim. Cosmochim. Acta* **72**, 1415 (2008).
- A. Morbidelli et al., *Meteorit. Planet. Sci.* **35**, 1309 (2000).
- D. P. O'Brien, A. Morbidelli, H. F. Levison, *Icarus* **184**, 39 (2006).
- M. J. Drake, K. Righter, *Nature* **416**, 39 (2002).
- We thank M. Norman and T. Grove for supplying the samples from Hawaii and Mount Shasta, respectively. This work was supported by NSF grant EAR-0229417 and the Carnegie Institution of Washington. M.S. also acknowledges the British Science and Technology Facility Council.

Supporting Online Material

www.sciencemag.org/cgi/content/full/328/5980/884/DC1

SOM Text

Figs. S1 to S3

Tables S1 and S2

References

22 December 2009; accepted 13 April 2010

10.1126/science.1186239

Narrow Primary Feather Rachises in *Confuciusornis* and *Archaeopteryx* Suggest Poor Flight Ability

Robert L. Nudds^{1*} and Gareth J. Dyke²

The fossil birds *Archaeopteryx* and *Confuciusornis* had feathered wings resembling those of living birds, but their flight capabilities remain uncertain. Analysis of the rachises of their primary feathers shows that the rachises were much thinner and weaker than those of modern birds, and thus the birds were not capable of flight. Only if the primary feather rachises were solid in cross-section (the strongest structural configuration), and not hollow as in living birds, would flight have been possible. Hence, if *Archaeopteryx* and *Confuciusornis* were flapping flyers, they must have had a feather structure that was fundamentally different from that of living birds. Alternatively, if they were only gliders, then the flapping wing stroke must have appeared after the divergence of *Confuciusornis*, likely within the enantiornithine or ornithurine radiations.

The earliest birds, including *Archaeopteryx* [Late Jurassic; 145 million years ago (Ma)] and *Confuciusornis* (Early Cretaceous; 120 Ma), have feathered wings that appear capable of generating some aerodynamic force (1). The likely magnitude of these forces and allied flight capabilities, however, remain controversial

(2–10). Some studies suggest that *Archaeopteryx* was capable of flapping flight (2, 3, 5, 9, 10); others have concluded the opposite (4, 6–8). The flight capabilities of *Confuciusornis* (11) have received less attention, despite hundreds of fossils being known (12–15), many with complete primary feathers (12) (Fig. 1A). Primary feathers are fundamental to bird flight, contributing 40 to 51% of the total wingspan (16), and they must be capable of sustaining lift forces without breaking (17). Sustained flight, in which a horizontal position is maintained in the air, requires that the flight surfaces of the bird support a lift force

¹Faculty of Life Sciences, University of Manchester, Manchester M13 9PT, UK. ²School of Biology and Environmental Science, University College Dublin, Belfield 4, Dublin, Ireland.

*To whom correspondence should be addressed. E-mail: robert.nudds@manchester.ac.uk

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

The following resources related to this article are available online at www.sciencemag.org (this information is current as of May 13, 2010):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:
<http://www.sciencemag.org/cgi/content/full/328/5980/887>

Supporting Online Material can be found at:
<http://www.sciencemag.org/cgi/content/full/328/5980/887/DC1>

This article **cites 17 articles**, 3 of which can be accessed for free:
<http://www.sciencemag.org/cgi/content/full/328/5980/887#otherarticles>

This article appears in the following **subject collections**:
Paleontology
<http://www.sciencemag.org/cgi/collection/paleo>

equal to body weight [mass × acceleration of gravity (*g*)]. Much higher forces, far in excess of body weight, are generated in extreme maneuvers and accelerations (18). Previous work indicates that feathers fail by buckling (17). Accordingly, we used Euler-Bernoulli beam theory to determine the load-bearing capabilities of the primary feathers of *Archaeopteryx* and *Confuciusornis* and infer their likely flight capabilities (19).

The mean first primary feather length of five *Confuciusornis* specimens was 207 ± 17 mm, and the average rachis diameter was 1.06 ± 0.12 mm, compared with 129 and 0.75 mm, respectively, for the longest primary of the Munich *Archaeopteryx* (Table 1), which is consistent with measurements by Elżanowski (20). There was no evidence of flattening during preservation of the fossil feathers, but flattening would have increased apparent diameters and therefore increased our estimates of feather strength. The primary feather lengths of

Confuciusornis and *Archaeopteryx* are close to those predicted for extant birds of similar size (Fig. 1B). In both fossil birds, however, the rachises are narrow for the birds' body size and feather lengths (Fig. 1, B and C). In contrast to the fossil birds, the feather morphologies of two similar-sized living birds (*Larus ridibundus* and *Columba palumbus*) and two larger soaring birds (*Diomedea exulans* and *Gyps fulvus*) are comparable to those predicted from the relationships for other extant birds determined by Worcester (21).

It is a well-established concept that the larger of two cylinders with walls of the same thickness relative to overall radius can withstand a larger moment before buckling (Fig. 2). The significant result here is that the moment required to buckle a *Confuciusornis* primary feather (0.0036 N·m) is two orders of magnitude less (Fig. 2) than that required (0.3234 N·m) if its diameter matched that of a primary from an extant bird with similar

feather length (4.8 mm). Similarly, the primary feathers of *Archaeopteryx* were one order of magnitude more likely to fail by buckling (critical moment = 0.0013 N·m versus 0.0652 N·m) than are the feathers of a similar-sized extant analog (Fig. 2). Even if the primary feathers are modeled as the strongest possible structure—a solid cross-section of keratin—the critical failure moment (tensile failure) is one order of magnitude below that expected for a similar-sized extant bird for both *Confuciusornis* (0.0264 N·m) and *Archaeopteryx* (0.0094 N·m).

The lift on the primary feathers (*L*_{primaries}) as a percentage of lift on the entire wing was calculated as 42.9% for *Confuciusornis* and 30.9% for *Archaeopteryx*. For comparison, the lift is 39.2% for *C. palumbus*, 29.8% for *L. ridibundus*, 11.1% for *D. exulans*, and 25.5% for *G. fulvus*. For lift generation equal to body weight, *L*_{primaries} forces must therefore be 1.05 N (*Confuciusornis*),

Fig. 1. (A) One of the specimens of *Confuciusornis sanctus* (BSP 1999 I 15) included in the study. The approximate length of one functional primary feather is marked with a black line. Scale bar is 20 mm. **(B)** Plots of rachis diameter (bottom) and feather length (top). The regression lines [from (21)] describe the relationships for extant birds (length versus body mass, $y = 0.21x^{0.30}$ (0.23 to 0.37) and diameter versus body mass $y = 0.004x^{0.37}$ (0.31 to 0.43)). **(C)** Plot of rachis diameter against feather length of extant birds. The regression line describing the relationship for modern birds is $y = 0.03x^{1.16}$ (0.99 to 1.33) (21). Data for *Archaeopteryx* are represented by dark gray diamonds and for *Confuciusornis* by light gray diamonds. Extant species are denoted by black circles. The order (left to right) in (B) is *L. ridibundus*, *C. palumbus*, *G. fulvus*, and *D. exulans*. In (C), the order is *C. palumbus*, *L. ridibundus*, *D. exulans*, and *G. fulvus*. Dotted lines are 95% confidence limits.

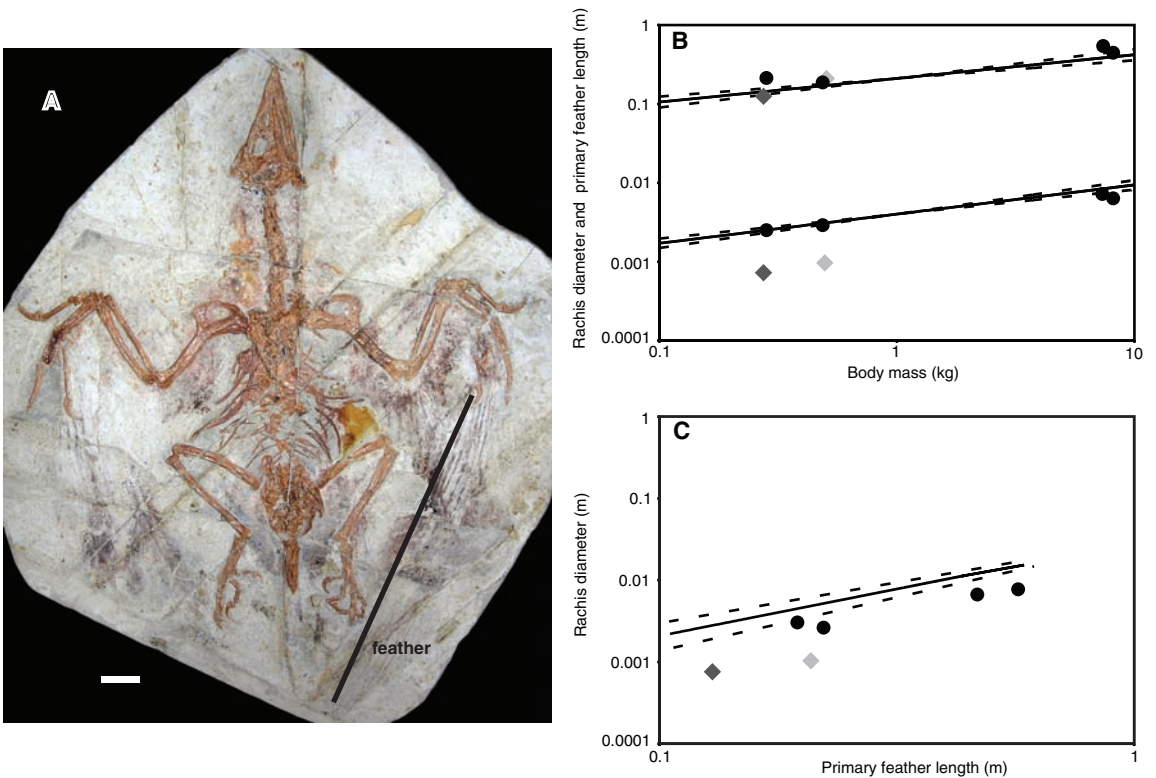


Table 1. The morphology, rachis buckling failure moments, and calculated lift forces for the species used in the study.

Species	Body mass (kg)	Wing length (m)	Rachis diameter (mm)	Primary length (mm)	Rachis buckling failure moment (N·m)	Moment exerted upon a single primary* (N·m)
<i>Confuciusornis</i>	0.500	0.350	1.06	207	0.0036	0.0092
<i>Archaeopteryx</i>	0.276	0.264	0.75	129	0.0013	0.0023
<i>C. palumbus</i>	0.490	0.345	2.97	194	0.0786	0.0077
<i>L. ridibundus</i>	0.284	0.460	2.60	220	0.0527	0.0039
<i>D. exulans</i>	8.190	1.622	6.60	460	0.8622	0.0871
<i>G. fulvus</i>	7.436	1.250	7.65	560	1.3426	0.2209

*Moment exerted by a lift force equal to body weight (mass × *g*).

0.42 N (*Archaeopteryx*), 0.94 N (*C. palumbus*), 0.42 N (*L. ridibundus*), 4.46 N (*D. exulans*), and 9.30 N (*G. fulvus*). For the two fossil birds, these forces exceed what the primaries could withstand, if they were of comparable structure to that of modern bird feathers (Fig. 3). The estimated buckling failure moment of their rachises was only 39% (*Confuciusornis*) and 55% (*Archaeopteryx*) of that generated upon the feather by a lift force equal to body weight (Fig. 3). The estimated rachis safety factors (that is, buckling failure force/lift force equal to body weight) of 13.6, 10.1, 9.9, and 6.1 for *L. ridibundus*, *C. palumbus*, *D. exulans*, and *G. fulvus*, respectively, were comparable to the previously estimated range for *C. livia* during routine flight [values of 9 to 17 (17)], confirming that these species are capable of flapping flight (Fig. 3) and, hence, validating our methodology. With feathers structurally similar to those of modern birds, *Confuciusornis* and *Archaeopteryx* could only have parachuted with their wings held dorsally, reducing the forces acting on the primaries while providing drag to reduce descent speed. A parachuting arrangement seems incongruent with their overall wing morphologies. Even when they are modeled as solid keratin beams, however, the primary feathers of the fossil birds only provide safety factors of 2.9 (*Confuciusornis*) and 4 (*Archaeopteryx*), which is well below those of extant birds (Fig. 3).

At each stage of the calculations, the parameter estimates were selected to provide the likely upper bounds of lift-supporting capabilities. For example, the force acting upon the distal part of the wing was divided equally between 10 feath-

ers. Therefore, we likely underestimated the load on the leading primaries, because not all primaries bear distal lift forces equally. Primary feathers 1 to 4 generally constitute the wing tip, whereas primary feathers numbered 5 and above commonly do not extend right to the tip of the wing, and their long axes are progressively oriented more dorsally. In addition, the chordwise lift distribution is generally at least twice as high cranially than caudally. The estimates of body mass are crucial to the calculations of lift force. For their feathers (if morphologically similar to those of modern birds) to be strong enough to sustain a lift force equal to body weight, *Confuciusornis* and *Archaeopteryx* would have had to weigh 0.215 and 0.188 kg, respectively, and their body masses are unlikely to be this low. Bird primary feathers have longitudinal furrows, and these are seen in *Archaeopteryx* (5). Because of the intractable nature of incorporating a furrow into the cross-section, however, the shaft was modeled without a furrow. Previous work (17) showed that the compressive stress required to buckle pigeon primaries, estimated using Euler-Bernoulli beam theory, was 21% higher than actually required. Therefore, the theoretical estimates here are probably high, and the feathers are actually structurally weaker.

Because even rudimentary forelimb movements can generate useful thrust (1, 22, 23), some thrust generation by these fossil birds cannot be discounted, but the vigorous flapping flight of modern birds is highly unlikely. Flight toward the gliding end of the ability spectrum is consistent with *Archaeopteryx*'s known musculature (4),

shoulder anatomy (6), feather vane anatomy, and feather shaft curvature (2). It is also congruent with the enlarged and excavated deltopectoral crest of *Confuciusornis* (12). Our results suggest that superficial resemblance of Mesozoic bird wing feathers to those of living birds should not necessarily be taken as an indication of the ability to withstand similar aerodynamic forces. Unless Mesozoic primary feathers were structurally different from those of extant birds, further refinements were required after the appearance of primary feathers for the evolution of a powered flapping flight stroke. Consequently, the origins of the modern wing-beat may have occurred among taxa that diverged subsequently to *Confuciusornis*, likely within the diverse Cretaceous enantiornithine or ornithurine radiations.

References and Notes

1. R. L. Nudds, G. J. Dyke, *Evolution* **63**, 994 (2009).
2. R. A. Norberg, in *The Beginnings of Birds: Proceedings of the International Archaeopteryx Conference*, Eichstätt 1984, M. K. Hecht, J. H. Ostrom, G. Viohl, P. Wellnhofer, Eds. (Freunde des Jura-Museums, Eichstätt, Germany, 1985), pp. 303–318.
3. S. L. Olson, A. Feduccia, *Nature* **278**, 247 (1979).
4. J. H. Ostrom, *Q. Rev. Biol.* **49**, 27 (1974).
5. S. Rietschel, in *The Beginnings of Birds: Proceedings of the International Archaeopteryx Conference*, Eichstätt 1984, M. K. Hecht, J. H. Ostrom, G. Viohl, P. Wellnhofer, Eds. (Freunde des Jura-Museums, Eichstätt, Germany, 1985), pp. 251–260.
6. P. Senter, *Acta Palaeontol. Pol.* **51**, 305 (2006).
7. J. R. Speakman, S. C. Thomson, *Nature* **370**, 514 (1994).
8. R. J. Vazquez, *J. Morphol.* **211**, 259 (1992).
9. P. Wellnhofer, in *Florida Symposium on Dinosaur-Bird Evolution*, P. J. Currie, E. B. Koppelhus, M. A. Shugar, Eds. (Indiana University Press, Fort Lauderdale, FL, 2004), pp. 282–300.
10. D. W. Walden, *Ibis* **113**, 349 (1971).
11. L. Hou, L. M. Chiappe, F. Zhang, C.-M. Chuong, *Naturwissenschaften* **91**, 22 (2004).
12. L. M. Chiappe, S. Ji, Q. Ji, M. A. Norell, *Bull. Am. Mus. Nat. Hist.* **242**, 1 (1999).
13. L. M. Chiappe, J. Marugán-Lobón, S. Ji, Z. Zhou, *Biol. Lett.* **4**, 719 (2008).
14. T. M. R. Fountaine, M. J. Benton, G. J. Dyke, R. L. Nudds, *Proc. Biol. Sci.* **272**, 289 (2005).
15. Z. Zhou, F. Zhang, *Acta Zool. Sin.* **50**, 913 (2004).
16. R. L. Nudds, *J. Avian Biol.* **38**, 515 (2007).
17. W. R. Corning, A. A. Biewener, *J. Exp. Biol.* **201**, 3057 (1998).
18. U. M. Norberg, *Vertebrate Flight: Mechanics, Physiology, Morphology, Ecology and Evolution* (Springer-Verlag, Berlin, 1990).
19. Materials and methods are available as supporting material on Science Online.
20. A. Elzanowski, in *Mesozoic Birds: Above the Heads of Dinosaurs*, L. M. Chiappe, L. M. Witmer, Eds. (Univ. of California Press, Berkeley, 2002), pp. 129–159.
21. S. E. Worchester, *J. Zool. (Lond.)* **239**, 609 (1996).
22. U. M. Norberg, *Am. Nat.* **126**, 303 (1985).
23. P. Burgers, L. M. Chiappe, *Nature* **399**, 60 (1999).
24. We thank J. Gardiner, G. Kaiser, J. Sigwart, and three anonymous referees. Collections access was kindly provided by O. Rauhut [Bayerische Staatssammlung fuer Palaeontologie und Geologie (BSP)], R. Brocke [Senckenberg Forschungsinstitut und Naturmuseum (SMF)], and J. Oelkers-Schaefer (SMF).

Supporting Online Material

www.sciencemag.org/cgi/content/full/328/5980/887/DC1
Materials and Methods
References and Notes

2 December 2009; accepted 12 April 2010
10.1126/science.1188895

Fig. 2. Graph of moment arm (N·m) against rachis diameter (m) [from equation 4 in (17)]. Dark gray diamonds represent *Archaeopteryx*, light gray diamonds represent *Confuciusornis*, and the black circles indicate the data for the four extant species (from left to right: *L. ridibundus*, *C. palumbus*, *D. exulans*, and *G. fulvus*). Black-bordered diamonds are predicted moments if the rachis diameter relative to feather length was the same as that predicted from extant birds. The line ($y = 0.006x^3$) describes the theoretical critical buckling moment as a function of thin-walled cylinder diameter.

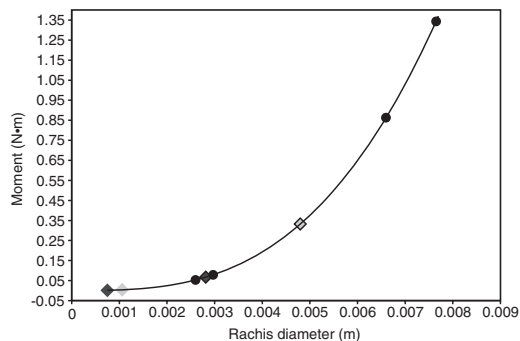
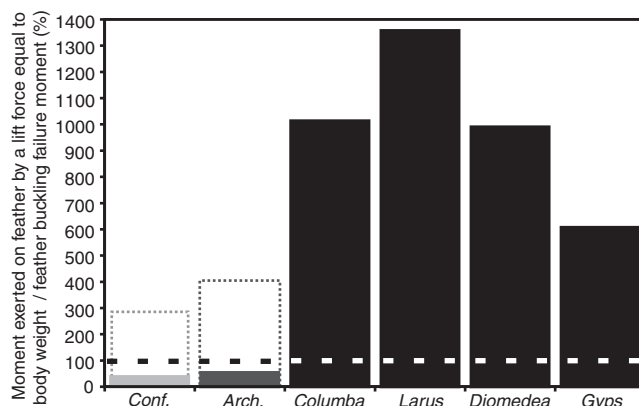


Fig. 3. Bar chart showing the critical buckling failure moment (solid bars) and bending tensile failure moment (dotted bordered open bars) of the primary feathers as a percentage of the moment that would be exerted upon the feather by a lift force equal to body mass times gravity. Dashed line represents a safety factor of 1.



This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

The following resources related to this article are available online at www.sciencemag.org (this information is current as of May 13, 2010):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/328/5980/890>

Supporting Online Material can be found at:

<http://www.sciencemag.org/cgi/content/full/science.1188919/DC1>

This article **cites 32 articles**, 14 of which can be accessed for free:

<http://www.sciencemag.org/cgi/content/full/328/5980/890#otherarticles>

This article has been **cited by** 1 articles hosted by HighWire Press; see:

<http://www.sciencemag.org/cgi/content/full/328/5980/890#otherarticles>

This article appears in the following **subject collections**:

Biochemistry

<http://www.sciencemag.org/cgi/collection/biochem>

Reconstitution of Outer Membrane Protein Assembly from Purified Components

Christine L. Hagan,^{1*} Seokhee Kim,^{1*} Daniel Kahne^{1,2†}

β -barrel membrane proteins in Gram-negative bacteria, mitochondria, and chloroplasts are assembled by highly conserved multi-protein complexes. The mechanism by which these molecular machines fold and insert their substrates is poorly understood. It has not been possible to dissect the folding and insertion pathway because the process has not been reproduced in a biochemical system. We purified the components that fold and insert *Escherichia coli* outer membrane proteins and reconstituted β -barrel protein assembly in proteoliposomes using the enzymatic activity of a protein substrate to report on its folding state. The assembly of this protein occurred without an energy source but required a soluble chaperone in addition to the multi-protein assembly complex.

The outer membranes of Gram-negative bacteria and the mitochondria and chloroplasts of higher eukaryotes contain proteins with β -barrel structure, which are assembled in their respective membranes by multi-protein machines (1–6). The folding and insertion of these β -barrels must be coordinated because they would have many unsatisfied hydrogen bonds in the membrane if they were inserted in an unfolded state, but conversely, they would be “inside out” if they folded first in the aqueous environment and were subsequently inserted. In order to understand how β -barrel proteins assemble into membranes, we purified the proteins comprising the *Escherichia coli* outer membrane protein (OMP)–folding machinery and established a reconstituted system to monitor the activity of this machinery.

The β -barrel assembly machine (Bam) in *E. coli* consists of an integral β -barrel protein, BamA (formerly YaeT), and four lipoproteins, BamB, -C, -D, and -E (formerly YfgL, NlpB, YfiO, and SmpA, respectively). Only BamA and BamD are essential for cell survival, but deleting or depleting any member of the complex causes defects in OMP assembly (6–9). BamA has homologs in prokaryotes and eukaryotes and contains five periplasmic polypeptide transport–associated (POTRA) domains, which scaffold the Bam lipoproteins (2–5, 10). Unfolded OMPs are delivered to this complex after their synthesis in the cytoplasm and translocation across the inner membrane by the secretion machinery (SecYEG) (Fig. 1A) (11). The chaperones, SurA, Skp, and DegP, prevent unfolded OMPs that are released from the Sec machine from aggregating and misfolding as they transit the periplasm. These chaperones are thought to transport OMPs in two parallel but separate pathways: one that relies on SurA and one that involves both Skp and DegP (12, 13). Most OMPs can be handled by either pathway, but SurA delivers the bulk of OMPs to the OM (14).

The process of OMP assembly can be monitored in isolated mitochondria (15) and in vivo in *E. coli* (16). We sought to develop an in vitro system to study the function of the Bam proteins. Expressing all five *bam* genes in a single strain produced a mixture of complexes and subcomplexes that could not be easily separated. Previous lipoprotein and POTRA domain–deletion experiments indicated that BamA binds BamB independently from BamCDE (8, 10); thus,

we expressed the subcomplexes BamAB and BamCDE separately and reconstituted the full complex in vitro (17). The reconstructed complex was identical to the native complex, as analyzed with blue native polyacrylamide gel electrophoresis (BN-PAGE) (Fig. 1B). The purified Bam complex had an apparent molecular weight that is too small to accommodate more than one copy of BamA, and the ratio of BamA:B:C:D was 1:1:1:1 (fig. S1). Either one or two copies of BamE may have been present; its small size prevents a definitive conclusion. The Bam complex may exist as a higher-order oligomer in the membrane in vivo, but the relative stoichiometry is expected to be as we have determined.

Purified Bam complex was incorporated into liposomes, as others have done to reconstitute complexes that handle membrane proteins (18–22). To follow OMP assembly in the proteoliposomes, we used a substrate OMP that possesses enzymatic activity, OmpT. OmpT is a β -barrel outer membrane protease that cleaves peptides between two consecutive basic residues, and its activity can be monitored by using a fluorogenic peptide (23). Urea-denatured OmpT was incubated with the periplasmic chaperone

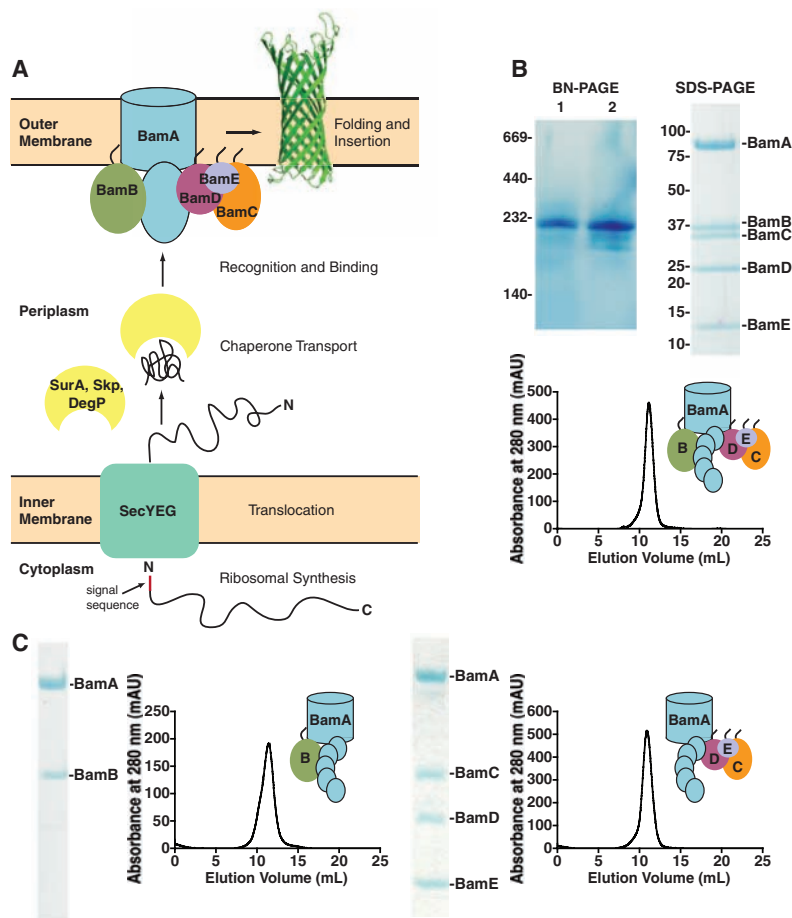


Fig. 1. The OMP assembly pathway and the purified Bam complex and subcomplexes. **(A)** The *E. coli* OMP biogenesis pathway. **(B)** Blue native gel analysis, SDS-PAGE, and gel filtration chromatogram of the purified Bam complex. The complex isolated from cells expressing the complex at basal levels (lane 1) and the overproduced and reconstructed complex (lane 2). **(C)** SDS-PAGE and gel filtration chromatograms of the purified two- and four-protein Bam subcomplexes.

¹Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA. ²Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA.

*These authors contributed equally to this work.

†To whom correspondence should be addressed. E-mail: kahne@chemistry.harvard.edu

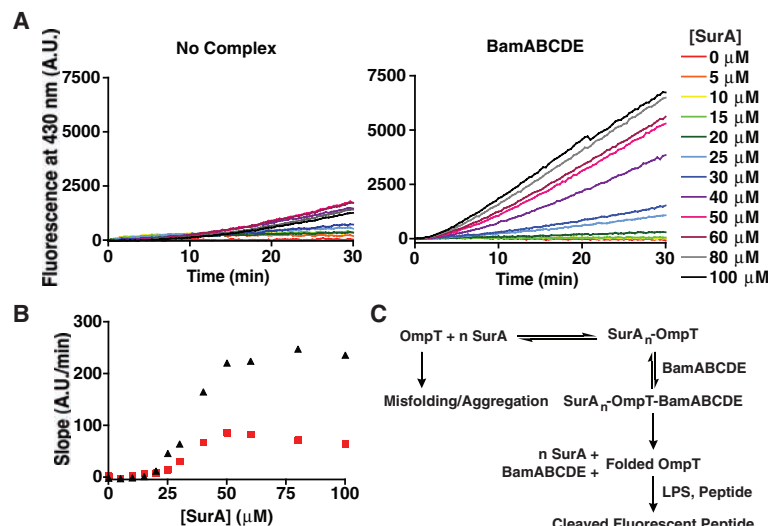


Fig. 2. SurA and the Bam complex facilitate OmpT assembly in proteoliposomes. **(A)** Preincubated solutions of urea-denatured OmpT with SurA were diluted [at time (t) = 0 min] into (left) liposomes or (right) proteoliposomes that contain the Bam complex. OmpT was present in each reaction at a final concentration of 10 μ M, and the concentration of SurA was varied from 0 to 100 μ M. **(B)** The amount of active OmpT (reflected in the slope from 15 to 30 min) in the reactions in (A) as a function of SurA concentration in the presence (black triangles) or absence (red squares) of the Bam complex. **(C)** Schematic of the reconstitution reaction pathway.

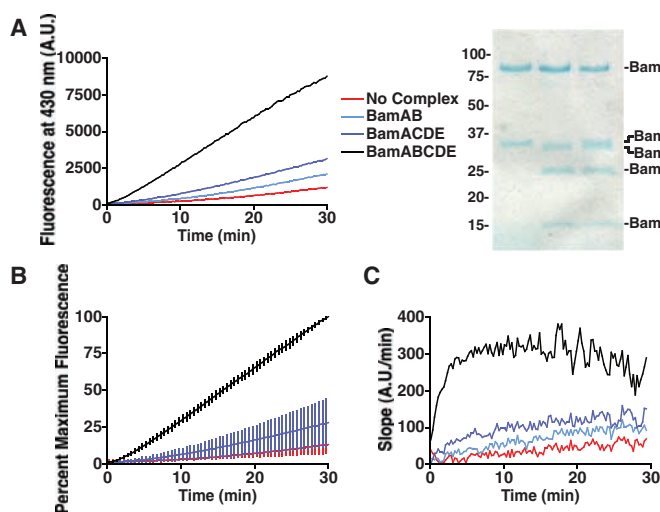
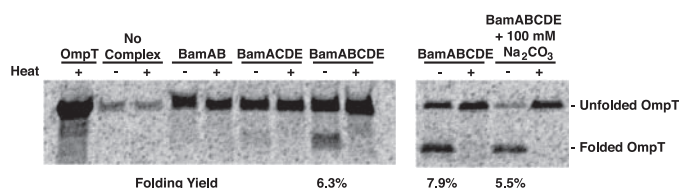


Fig. 3. OmpT assembly requires specific components of the Bam complex. **(A)** Fluorescence produced in the presence of the Bam complex and subcomplexes (left). OmpT and SurA are diluted to final concentrations of 10 μ M and 100 μ M, respectively. SDS-PAGE of these proteoliposomes indicates that they contain equal amounts of the complexes (right). **(B)** Eight different experiments were normalized to their maximum fluorescence values and then averaged. The error bars represent the SD among these experiments. **(C)** The gradient of the fluorescence data in (A).

Fig. 4. OmpT assembled by the Bam complex is folded on SDS-PAGE and resistant to membrane extraction. Shown are pellets of the reactions of SurA and 35 S-labeled OmpT with proteoliposomes containing the Bam complexes. 35 S-labeled OmpT and SurA were diluted to final concentrations of 0.4 μ M and 100 μ M, respectively. Extraction in 100 mM sodium carbonate results in a threefold enrichment of the folded material relative to the unfolded material.



SurA, and this chaperone-OmpT complex was diluted into solutions containing the proteoliposomes, the fluorogenic peptide, and lipopolysaccharide (LPS), which is required for OmpT activity (17). OmpT assembly was observed as an increase in the rate of fluorescence production.

OmpT activity increased with the concentration of SurA and in the presence of the Bam complex (Fig. 2A and figs. S2 and S4). OmpT activity saturated at the same SurA concentration in the presence or absence of the complex, which suggests that SurA plays a general role in delivering the unfolded protein in a state that can be folded and also that the Bam complex handles OmpT bound to SurA (Fig. 2, B and C). The concentrations of SurA at which we saw a substantial improvement in OmpT assembly (above 20 μ M) are comparable with the reported binding constants for SurA and model peptides (1 to 14 μ M) (24, 25). OmpT contains several putative SurA-binding sites (26, 27), and the sigmoidal relationship may indicate that multiple SurA molecules are involved in generating a SurA_n-OmpT complex that can be folded efficiently.

We then wanted to determine whether the assembly observed in the presence of the Bam complex reflected the function of the complex and not a nonspecific property of the proteoliposomes (28, 29). We compared the activity of the five-protein Bam complex with subcomplexes containing two (BamAB) and four proteins (BamACDE), which had comparable stability as judged by their behavior on gel filtration chromatography (Fig. 1C). The components of these subcomplexes were incorporated into proteoliposomes in equal proportion to those of the five-protein complex (Fig. 3A). Cleavage of the fluorogenic substrate was greatly reduced in proteoliposomes containing the subcomplexes as compared with those containing the five-protein complex (Fig. 3A and fig. S3). The activity of the four-protein complex was always less than that of the five-protein complex (Fig. 3B). In the first five minutes of the experiment, OmpT assembly was orders of magnitude faster in the presence of the five-protein complex (Fig. 3C). These large differences in activity at early time points are relevant given that *in vivo* pulse-chase experiments indicate that OMP assembly occurs in the cell on a similar time scale of 30 s to several minutes (30, 31). The four-protein complex appeared to have some activity on this time scale, which BamB dramatically improved. Although there was a background rate of OmpT folding in the absence of the Bam complex, the Bam proteins significantly improved the kinetics of OmpT assembly.

In order to verify that increased fluorescence correlated with increased amounts of folded OmpT, we ran the products of our folding reactions on SDS-PAGE gels without and with prior heat denaturation. β -barrel proteins remain folded on SDS-PAGE if they are not boiled and consequently run faster than their unfolded forms. Purified 35 S-labeled OmpT was incubated with SurA and then diluted into proteoliposomes, as in the fluorescence experiments. After 30 min, the

reaction solutions were centrifuged and the pellets were run on SDS-PAGE (Fig. 4 and fig. S5). Much more folded OmpT was observed in the reaction of SurA₉-OmpT with the full Bam complex than in the reactions with the subcomplexes. The yield of folded protein represented by the faster migrating band was about 7%, as determined by five separate experiments. The folded OmpT was resistant to extraction from the pellet by incubation in 100 mM sodium carbonate for 30 min, suggesting that the folded protein had been integrated into the membrane.

The nonessential lipoprotein BamB is important in the assembly of OMP substrates delivered by SurA; the five-protein complex had significantly higher activity than the four-protein complex, BamACDE. Thus, some or all of the BamCDE proteins are required for OmpT assembly, but BamB appears to have a specific function that, when coordinated with the other components, increases the activity of the complex. The functional importance of both SurA and BamB is consistent with *in vivo* observations that these proteins play related but not redundant roles in OMP biogenesis. Deleting *surA* or *bamB* produces identical phenotypes with respect to the kinetics of conversion of unfolded mature LamB to folded monomers (31). In addition, simultaneous deletion of the *surA* and *bamB* genes results in a phenotype that is severely defective in the assembly of OMPs and much sicker than either single-deletion strain (32, 33). Our results also suggest that SurA and BamB both affect the efficiency of OMP assembly. Our highly simplified system thus appears to recapitulate elements of the cellular process and clearly demonstrates that a nonessential protein can alter the activity of essential proteins to regulate the efficiency of the process that they catalyze.

We do not know if the Bam complex in proteoliposomes can process multiple substrate molecules, but clearly the Bam complex can assemble a β -barrel protein in the membrane without an input of energy. In contrast, the inner membrane secretion machinery couples energy-driven translocation to insertion (11). Although it remains possible that β -barrel assembly may be coupled to an energy-driven process in the cell, it is remarkable that these six proteins (the Bam complex and a chaperone) are sufficient to perform the folding and insertion process. Together, they perform a chemical transformation that we do not understand, and that they are able to do so without energy suggests that their structure inherently facilitates OMP assembly.

References and Notes

- W. C. Wimley, *Curr. Opin. Struct. Biol.* **13**, 404 (2003).
- S. Reumann, J. Davila-Aponte, K. Keegstra, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 784 (1999).
- R. Voulhoux, M. P. Bos, J. Geurtsen, M. Mols, J. Tommassen, *Science* **299**, 262 (2003).
- N. Wiedemann et al., *Nature* **424**, 565 (2003).
- S. A. Paschen et al., *Nature* **426**, 862 (2003).
- T. Wu et al., *Cell* **121**, 235 (2005).
- U. S. Eggert et al., *Science* **294**, 361 (2001).
- J. C. Malinverni et al., *Mol. Microbiol.* **61**, 151 (2006).
- J. G. Sklar et al., *Proc. Natl. Acad. Sci. U.S.A.* **104**, 6400 (2007).

- S. Kim et al., *Science* **317**, 961 (2007).
- A. J. Driessen, N. Nouwen, *Annu. Rev. Biochem.* **77**, 643 (2008).
- A. E. Rizzitello, J. R. Harper, T. J. Silhavy, *J. Bacteriol.* **183**, 6794 (2001).
- J. G. Sklar, T. Wu, D. Kahne, T. J. Silhavy, *Genes Dev.* **21**, 2473 (2007).
- D. Vertommen, N. Ruiz, P. Leverrier, T. J. Silhavy, J. F. Collet, *Proteomics* **9**, 2432 (2009).
- S. Kutik et al., *Cell* **132**, 1011 (2008).
- R. Ieva, H. D. Bernstein, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 19120 (2009).
- Materials and methods are available as supporting material on Science Online.
- L. Brundage, J. P. Hendrick, E. Schiebel, A. J. Driessen, W. Wickner, *Cell* **62**, 649 (1990).
- C. V. Nicchitta, G. Blobel, *Cell* **60**, 259 (1990).
- J. Akimaru, S. Matsuyama, H. Tokuda, S. Mizushima, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 6545 (1991).
- D. Görlich, T. A. Rapoport, *Cell* **75**, 615 (1993).
- T. Yakushi, K. Masuda, S. Narita, S. Matsuyama, H. Tokuda, *Nat. Cell Biol.* **2**, 212 (2000).
- R. A. Kramer, D. Zandwijken, M. R. Egmond, N. Dekker, *Eur. J. Biochem.* **267**, 885 (2000).
- E. Bitto, D. B. McKay, *J. Biol. Chem.* **278**, 49316 (2003).
- G. Hennecke, J. Nolte, R. Volkmer-Engert, J. Schneider-Mergener, S. Behrens, *J. Biol. Chem.* **280**, 23540 (2005).
- X. Xu, S. Wang, Y. X. Hu, D. B. McKay, *J. Mol. Biol.* **373**, 367 (2007).
- J. E. Mogensen, D. E. Otzen, *Mol. Microbiol.* **57**, 326 (2005).
- J. H. Kleinschmidt, *Chem. Phys. Lipids* **141**, 30 (2006).
- N. K. Burgess, T. P. Dao, A. M. Stanley, K. G. Fleming, *J. Biol. Chem.* **283**, 26748 (2008).
- C. Jansen, M. Heutink, J. Tommassen, H. de Cock, *Eur. J. Biochem.* **267**, 3792 (2000).
- A. R. Ureta, R. G. Endres, N. S. Wingreen, T. J. Silhavy, *J. Bacteriol.* **189**, 446 (2007).
- C. Onufryk, M. L. Crouch, F. C. Fang, C. A. Gross, *J. Bacteriol.* **187**, 4552 (2005).
- N. Ruiz, B. Falcone, D. Kahne, T. J. Silhavy, *Cell* **121**, 307 (2005).
- The Biophysics Resource of the Keck Facility at Yale University performed the size exclusion–light scattering analysis and amino acid analysis. This work is supported by NIH grant AI081059. C.L.H. is supported by a NSF Graduate Research Fellowship and a Kellogg fellowship from Amherst College.

Supporting Online Material

www.sciencemag.org/cgi/content/full/science.1188919/DC1
Materials and Methods
Figure S1 to S6
Tables S1 and S2

1 March 2009; accepted 29 March 2009

Published online 8 April 2009;

10.1126/science.1188919

Include this information when citing this paper.

Additive Genetic Breeding Values Correlate with the Load of Partially Deleterious Mutations

Joseph L. Tomkins,^{1,*} Marissa A. Penrose,¹ Johan Greeff,^{1,2} Natasha R. LeBas¹

The mutation-selection–balance model predicts most additive genetic variation to arise from numerous mildly deleterious mutations of small effect. Correspondingly, “good genes” models of sexual selection and recent models for the evolution of sex are built on the assumption that mutational loads and breeding values for fitness-related traits are correlated. In support of this concept, inbreeding depression was negatively genetically correlated with breeding values for traits under natural and sexual selection in the weevil *Callosobruchus maculatus*. The correlations were stronger in males and strongest for condition. These results confirm the role of existing, partially recessive mutations in maintaining additive genetic variation in outbred populations, reveal the nature of good genes under sexual selection, and show how sexual selection can offset the cost of sex.

Partially recessive, deleterious mutations of small effect are thought to be responsible for much of the observed additive genetic variation in traits that are closely associated with fitness (1–4). These mutations are also central to “good genes” models of sexual selection, which posit that the additive genetic breeding values of males must correlate negatively with mutational load (5–8). The contribution of nonadditive allelic effects to additive genetic variation through mutation-selection balance has empirical support from mutation-accumulation experiments (9) and molecular-evolution surveys (10), but not from recent biometric analyses (11–13).

¹Centre for Evolutionary Biology, School of Animal Biology, The University of Western Australia, Nedlands, WA 6009, Australia. ²Department of Agriculture and Food, 3 Baron-Hay Court, South Perth, WA 6151, Australia.

*To whom correspondence should be addressed. E-mail: jtomkins@cyllene.uwa.edu.au

Here we develop Fox’s proposal (14) for detecting among-family variation in inbreeding depression. Inbreeding depression is the reduction in fitness observed in the offspring of related (compared with unrelated) parents (4, 15) and is believed to be predominantly caused by the increased homozygosity of numerous, at least partially recessive, deleterious mutations of small effect (4, 15). By estimating the correlation between the load of mutations carried by a pair of parents and their breeding value, we were able to test the central assumption of good genes models of sexual selection and contribute to our understanding of the maintenance of genetic variation.

We used a seven-generation pedigree of cow-pea weevils *Callosobruchus maculatus* (Coleoptera: Bruchidae) reared on black-eye beans (*Vigna unguiculata*) to estimate predicted breeding values, heritabilities, and inbreeding depression in a range of life-history (LH) and morphological

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

The following resources related to this article are available online at www.sciencemag.org (this information is current as of May 13, 2010):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/328/5980/892>

Supporting Online Material can be found at:

<http://www.sciencemag.org/cgi/content/full/328/5980/892/DC1>

This article **cites 25 articles**, 5 of which can be accessed for free:

<http://www.sciencemag.org/cgi/content/full/328/5980/892#otherarticles>

This article appears in the following **subject collections**:

Evolution

<http://www.sciencemag.org/cgi/collection/evolution>

(M) traits (16). Condition, or resource-acquisition ability, is a potential source of genetic variation for other LH traits (7, 8, 17), because resource acquisition represents a large mutational target, as virtually all genes can have effects on the ability to acquire resources, even if they are only subtle (7, 8, 17). We directly quantified resource-acquisition ability as the efficiency with which beetle larvae converted beans into beetle biomass (16). Large male size is sexually selected in *C. maculatus* [(18), also see supporting online material (SOM) text], and female size correlates with fecundity [(18, 19) and SOM text]. We documented significant heritable variation in all of the M traits we measured and some, but not all, of the LH traits (table S1). We also found that females were, in general, more susceptible to inbreeding depression than males (fig. S1) and, as in other studies, that this was particularly true for the longevity of virgin females (20–22).

Estimating among-family variation in inbreeding depression is often confounded by the effects of the genes of the out-crossed parent, as well as the genes shared with the related (or selfed) parent (4, 14, 16, 23). Thus, we adapted a method Fox (14) proposed in which, instead of comparing inbred and outbred offspring means, the outbred offspring trait values are predictions made on the basis of pedigree information, and only these are compared with the observed trait values of the inbred individuals (fig. S5) (16). In our experiment, the family-level estimate of inbreeding depression was correlated with the mid-parent breeding values of those families [i.e., the average best linear unbiased predictor of the brother and sister (fig. S6) (16)]. We used randomizations to estimate the null hypothesis for each correlation and subtracted this from the observed coefficient to obtain an estimate of the genetic correlation between inbreeding depression and breeding value (termed “ r_{ID-BV} ”) (16). Across all of the traits we studied, the average r_{ID-BV} coefficient was -0.24 (95% confidence interval: -0.33 to -0.15), significantly less than zero ($t_{16} = 5.69$, $P < 0.001$). Therefore, at the family level, greater inbreeding depression is associated with lower breeding values, whereas less inbreeding depression is associated with larger breeding values. The significant deviation from zero of the overall r_{ID-BV} coefficient suggests that, in this outbred population, the additive genetic breeding value for these traits is correlated with the load of mutations carried by the individual. The fact that the correlation is negative indicates that individuals with greater breeding values have a lower mutational load than individuals with lower breeding values. This negative genetic correlation supports good genes models of sexual selection (5–8) and is consistent with the predictions of mutation-selection–balance theory (4): Mutations that affect the breeding value of individuals within a population tend to be partially recessive (that is, they affect trait expression in the heterozygous state) but have greater effects when they are homozygous.

To further understand the patterns of variation in r_{ID-BV} coefficients, we analyzed genera-

tional variation in the correlation in relation to sex, type of trait (LH or M), and the trait nested within type. We found no significant interactions. However, there was significant variation across generations in r_{ID-BV} coefficients ($F_{4,58} = 2.91$, $P = 0.029$), suggesting that environmental conditions affect the correlation, which is to be expected because both heritability and inbreeding depression are influenced by the environment. There was no evidence for a consistent change in r_{ID-BV} coefficients with generation ($b = 0.013 \pm 0.019$) that might indicate selection due to the experimental design. Traits nested within LH and M did not vary significantly in the magnitude of r_{ID-BV} coefficients in LH or M ($F_{5,58} = 1.209$, $P = 0.316$), but there was a highly significant difference between LH and M trait types ($F_{1,58} = 48.43$, $P < 0.001$), with M traits having stronger correlations than LH traits.

Females showed stronger inbreeding depression than males for all traits in general, but particularly for longevity (SOM text). The latter is known to be due to a difference in the degree of dominance of alleles affecting longevity in males and females (22). With all things equal, such as heritability ($h^2_{\text{female}} = 0.31 \pm 0.06$, $h^2_{\text{male}} = 0.28 \pm 0.03$, paired $t_6 = 0.69$, $P = 0.5$), we might expect stronger r_{ID-BV} coefficients in females than in males. However, the sex effect we observed was significantly contrary to this expectation ($F_{1,58} = 4.642$, $P = 0.035$) (Fig. 1), with males showing stronger effects than females in both M (e.g., eclosion weight, efficiency) and LH traits (e.g., development time). This finding illustrates how, by revealing the effects that partially recessive alleles have on additive genetic variation, the r_{ID-BV} coefficient provides a different insight into genetic architecture than studies of inbreeding alone.

We quantified the ability of larval beetles to acquire resources from their environment by measuring the weight of beetle produced from the amount of bean that was eaten (16). Under-

standing how mutational load affects genetic variation in resource acquisition is important, because this variation is thought to be a source for genetic variation in other aspects of life history (8, 17). Consistent with this hypothesis, in our population, efficiency and eclosion weight showed the strongest negative r_{ID-BV} coefficient with marked differences between the sexes (Fig. 1).

The inbreeding in our experiment caused only one-quarter of the alleles to become homozygous by descent; this level of inbreeding explained 6 and 16% (i.e., r^2_{ID-BV}) of the variation in breeding values of LH and M traits, respectively. By extrapolation, if the whole genome were to be homozygous by descent (that is, if we multiply our r^2_{ID-BV} by 4), we would find that 24% (female = $23 \pm 9\%$, male = $26 \pm 9\%$) and 65% (female = $53 \pm 8\%$, male = $77 \pm 8\%$) of the variation in breeding values might be attributable to partially recessive deleterious mutations for these LH and M traits, respectively. Thus, extrapolating to a selfed cross suggests that partially recessive deleterious mutations might explain most, or possibly all, of the additive genetic variation in male condition and morphology (fig. S2). Nevertheless, not all of the additive genetic variation in these traits is explained through this extrapolation, suggesting that other forms of selection also have a role in maintaining genetic variation in these traits (11–13).

The good genes hypothesis of sexual selection assumes that the mutational load of males determines their reproductive success (5–8). Under this assumption, the two-fold cost of sex can be overcome (24, 25). Our data link the breeding values for male condition and traits under sexual selection with direct estimates of the load of mutations that males carry (r_{ID-BV} coefficients). We observed that male r_{ID-BV} coefficients generally exceeded that of females. Such elevated genotypic condition-dependence is thought to arise when costly traits are under selection (8) and

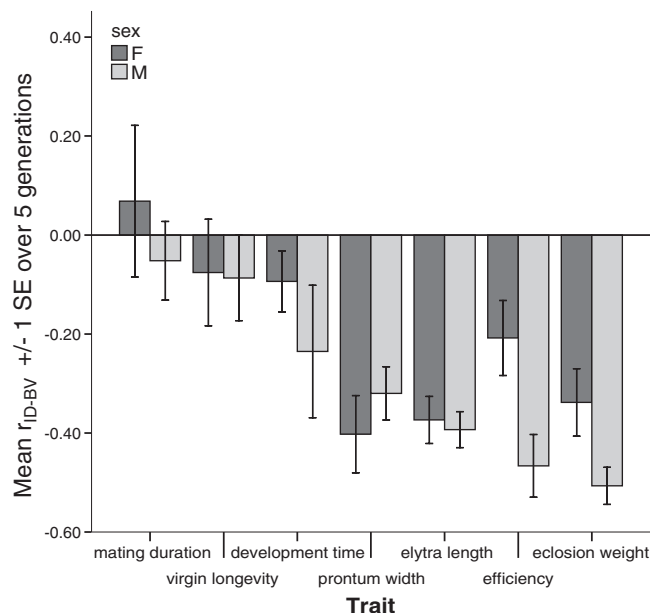


Fig. 1. Mean genetic correlation between inbreeding depression and breeding value for life-history and morphological traits in males and females. Error bars indicate 1 SE.

could occur in *C. maculatus* through sexual selection on males (18, 26–28). If sexual selection is responsible for the greater strength of the r_{ID-BV} coefficients in males, it raises the possibility of positive feedback, where sexual selection increases the contribution of deleterious mutations to trait expression, in turn increasing both good genes benefits from sexual selection and the benefit of sex itself.

References and Notes

- N. H. Barton, P. D. Keightley, *Nat. Rev. Genet.* **3**, 11 (2002).
- B. Charlesworth, K. A. Hughes, in *Evolutionary Genetics: From Molecules to Morphology*, R. S. Singh, C. B. Crimbas, Eds. (Cambridge Univ. Press, Cambridge, 1999), pp. 369–392.
- M. Lynch, L. Latta, J. Hicks, M. Giorgianni, *Evolution* **52**, 727 (1998).
- M. Lynch, B. Walsh, *Genetics and Analysis of Quantitative Traits* (Sinauer, Sunderland, MA, 1998).
- Y. Iwasa, A. Pomiankowski, S. Nee, *Evolution* **45**, 1431 (1991).
- J. Radwan, *Genetica* **134**, 113 (2008).
- J. L. Tomkins, J. Radwan, J. S. Kotiaho, T. Tregenza, *Trends Ecol. Evol.* **19**, 323 (2004).
- L. Rowe, D. Houle, *Proc. Biol. Sci.* **263**, 1415 (1996).
- D. L. Halligan, P. D. Keightley, *Annu. Rev. Ecol. Syst.* **40**, 151 (2009).
- A. Eyre-Walker, P. D. Keightley, *Nature* **397**, 344 (1999).
- B. Charlesworth, T. Miyo, H. Borthwick, *Genet. Res.* **89**, 85 (2007).
- J. K. Kelly, J. H. Willis, *Evolution* **55**, 937 (2001).
- J. K. Kelly, *Genetics* **164**, 1071 (2003).
- C. W. Fox, *Am. J. Bot.* **92**, 1929 (2005).
- B. Charlesworth, J. H. Willis, *Nat. Rev. Genet.* **10**, 783 (2009).
- Materials and methods are available as supporting material on Science Online.
- D. Houle, *Evolution* **45**, 630 (1991).
- C. W. Fox, U. M. Savalli, *Ethol. Ecol. Evol.* **11**, 49 (1999).
- F. J. Messina, *Heredity* **71**, 623 (1993).
- C. W. Fox, M. E. Czesak, W. G. Wallin, *J. Evol. Biol.* **17**, 1007 (2004).
- C. W. Fox et al., *Genetics* **174**, 763 (2006).
- C. W. Fox, K. L. Scheibly, B. P. Smith, W. G. Wallin, *Bull. Entomol. Res.* **97**, 49 (2007).
- S. T. Schultz, J. H. Willis, *Genetics* **141**, 1209 (1995).
- A. F. Agrawal, *Nature* **411**, 692 (2001).
- S. Siller, *Nature* **411**, 689 (2001).
- P. E. Eady, *Behav. Ecol. Sociobiol.* **36**, 25 (1995).
- S. Paukku, J. S. Kotiaho, *J. Insect Physiol.* **51**, 1220 (2005).
- C. W. Fox, R. C. Stillwell, W. G. Wallin, L. J. Hitchcock, *Funct. Ecol.* **20**, 1003 (2006).
- We thank J. Chan for the maintenance of the pedigree and the collection and management of the data set; R. Black and F. Gonzalez for statistical advice; B. Booth, A. Sutton, and S. Jennings for their assistance with the experiment; and N. Colegrave, W. Hazel, M. Puurtinen, J. Radwan, M. Ritchie, and L. Simmons for their comments on the manuscript. This work was supported by Australian Research Council fellowships to J.L.T. and N.R.L.

Supporting Online Material

www.sciencemag.org/cgi/content/full/328/5980/892/DC1

Materials and Methods

SOM Text

Figs. S1 to S6

Tables S1 to S10

References

5 February 2010; accepted 24 March 2010

10.1126/science.1188013

Erosion of Lizard Diversity by Climate Change and Altered Thermal Niches

Barry Sinervo,^{1,15*} Fausto Méndez-de-la-Cruz,² Donald B. Miles,^{3,15} Benoit Heulin,⁴ Elizabeth Bastiaans,¹ Maricela Villagrán-Santa Cruz,⁵ Rafael Lara-Resendiz,² Norberto Martínez-Méndez,² Martha Lucía Calderón-Espinosa,⁶ Rubi Nelsi Meza-Lázaro,² Héctor Gadsden,⁷ Luciano Javier Avila,⁸ Mariana Morando,⁸ Ignacio J. De la Riva,⁹ Pedro Victoriano Sepulveda,¹⁰ Carlos Frederico Duarte Rocha,¹¹ Nora Ibargüengoytia,¹² César Aguilar Puntriano,¹³ Manuel Massot,¹⁴ Virginie Lepetz,^{15†} Tuula A. Oksanen,¹⁶ David G. Chapple,¹⁷ Aaron M. Bauer,¹⁸ William R. Branch,¹⁹ Jean Clobert,¹⁵ Jack W. Sites Jr.²⁰

It is predicted that climate change will cause species extinctions and distributional shifts in coming decades, but data to validate these predictions are relatively scarce. Here, we compare recent and historical surveys for 48 Mexican lizard species at 200 sites. Since 1975, 12% of local populations have gone extinct. We verified physiological models of extinction risk with observed local extinctions and extended projections worldwide. Since 1975, we estimate that 4% of local populations have gone extinct worldwide, but by 2080 local extinctions are projected to reach 39% worldwide, and species extinctions may reach 20%. Global extinction projections were validated with local extinctions observed from 1975 to 2009 for regional biotas on four other continents, suggesting that lizards have already crossed a threshold for extinctions caused by climate change.

Global climate change affects organisms in all biomes and ecosystems. Two natural compensatory responses are possible. Given enough time and dispersal, species may shift to more favorable thermal environments, or they may adjust to new environments by behavioral plasticity, physiological plasticity, or adaptation. Alternatively, failure to adjust or adapt culminates in demographic collapse and extinction. Despite accumulating evidence of contemporary climate change affecting species ranges and phenologies (1–3), evidence of extinctions at either local or global scales is lacking (4–6). Moreover, current forecasting models (7, 8) are not calibrated with actual extinctions, but are premised on hypothesized effects of thermal physiology on demography and extinction. Alternatively, models are based on range shifts or species-area relations in mobile species

(1), but not extinctions (9). Hence, there is still much uncertainty regarding the expected magnitude of extinctions resulting from climate change (10).

Empirical validation of global extinction forecasts requires three forms of evidence. First, actual extinctions should be linked to macroclimate and validated to biophysical thermal causes arising from microclimate (11). Second, the pace of climate change should compromise thermal adaptation (10), such that evolutionary rates lag behind global warming owing to constraints on thermal physiology (12, 13). Third, extinctions due to climate should be global in extent.

From 2006 to 2008, we resurveyed 48 *Sceloporus* lizard species at 200 sites in Mexico that were first sampled in 1975 to 1995, and 12% of sites were locally extinct by 2009 (table S1).

Although *Sceloporus* lizards are heliotherms that bask and require solar radiation to attain physiologically active body temperatures (T_b) (14, 15), activity in hot weather may result in T_b exceeding CT_{max} , the critical thermal maximum, leading to death. Lizards retreat to cool refuges rather than risk death by overheating. However, hours of restriction (h_r) in thermal refuges limit foraging, constraining costly metabolic functions like growth, maintenance, and reproduction, thereby undermining population growth rates and raising extinction risk. Lizards could evolve higher T_b , but this brings them closer to CT_{max} , which increases risk of overheating. Extinction risk may increase because of other thermal adaptations. For example, viviparity, which is posited to be a thermal adaptation to cold climates (16), may elevate extinction risk because high T_b can compromise embryonic development in utero (17).

We analyzed rate of change in maximum air temperature T_{max} at 99 Mexican weather stations and constructed climate surfaces (tables S2 and S3, 1973 to 2008; fig. S1). Rate of change in T_{max} was greatest for winter-spring (January to May; fig. S1 and table S3A) and increased faster in northern and central México and at high elevation, as evidenced by significant coefficients for fitted climate surfaces. We found a correlation between rate of change in T_{max} during winter-spring breeding periods and local extinctions of *Sceloporus* species (table S3).

Many viviparous species in México are confined to high-elevation “islands,” where climate change has been most rapid. Logistic regression and multiple regression with phylogenetic independent contrasts (18, 19) revealed that extinction risk was significantly related to low latitudinal and altitudinal range limits (Fig. 1, A and B), where thermal physiology and/or ecological interactions limit species (20, 21). Phylogenetic correlation analysis (18) showed that extinction

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

The following resources related to this article are available online at www.sciencemag.org (this information is current as of May 13, 2010):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/328/5980/894>

Supporting Online Material can be found at:

<http://www.sciencemag.org/cgi/content/full/328/5980/894/DC1>

This article **cites 28 articles**, 8 of which can be accessed for free:

<http://www.sciencemag.org/cgi/content/full/328/5980/894#otherarticles>

This article has been **cited by 1** articles hosted by HighWire Press; see:

<http://www.sciencemag.org/cgi/content/full/328/5980/894#otherarticles>

This article appears in the following **subject collections**:

Ecology

<http://www.sciencemag.org/cgi/collection/ecology>

risk of viviparous lizards (18%) was twice that of oviparous lizards (9%, $n = 10000$ bootstrap replications $P < 0.001$). Moreover, multiple regression based on phylogenetic independent contrasts (PICs; Fig. 1C and table S4) showed that extinction risk of viviparous taxa was significantly related to low T_b ($T_{b,viviparous} = 31.8^\circ\text{C} \pm 0.31$ [SE], $T_{b,oviparous} = 34.8^\circ\text{C} \pm 0.40$, $t = 5.92$, $P < 0.0001$) and cool montane habitats ($T_{air,viviparous} = 22.4^\circ\text{C} \pm 1.79$, $T_{air,oviparous} = 28.39^\circ\text{C} \pm 1.38$, $t = 2.89$, $P < 0.006$), where climate has changed most rapidly in México.

To validate patterns of extinction risk and T_b , we deployed thermal models (22) that record operative temperatures (T_e) at two extinct and two persistent Yucatán sites of *S. serrifer*. Hours of restriction in activity (h_r) during reproduction was significantly higher at extinct versus persistent sites ($t = 9.26$, $P < 0.0001$). By April 2009, h_r at extinct Yucatán sites had become so severe that if *S. serrifer* were still present, it would have to retreat shortly after emergence (fig. S4A). Daily T_{max} was positively correlated with h_r assessed by T_e ($P < 0.001$, fig. S4B). The relation between h_r as a function of T_{max} relative to *S. serrifer*'s T_b [$h_r = 6.12 + 0.74 \times (T_{max} - T_b)$,

eq. S2 (23)] is a general formula for predicting extinctions.

We modeled extinct/persistence status based on values for h_r at *Sceloporus* sites derived from eq. S2 (23). The Yucatán ground truth for *S. serrifer* suggests that extinction occurs when h_r exceeds 4. We calibrated this value with extinct/persistent *Sceloporus* sites. Goodness-of-fit tests of the model indicate that the best fit for observed and predicted extinctions at *Sceloporus* sites is $h_r > 3.85$. If a species with a given T_b at a given geo-referenced site, subjected to $T_{max,j}$, experienced $h_r > 3.85$ during the 2-month reproductive period (March to April), we assumed that it would go extinct by 2009. Association of predicted and observed extinctions from this physiological model was significant for oviparous ($\chi^2 = 49.0$, $P < 0.001$) and viviparous taxa ($\chi^2 = 4.2$, $P < 0.04$).

As demography of high-elevation taxa becomes compromised due to climate change, species at low elevation that were previously limited by physiology and competition should expand into historically cooler habitat that is now warmer (20, 24), perhaps accelerating extinction of high-elevation forms. For viviparous taxa, six erroneously assigned extinct sites involved six of the eight cases of range expansion by low-elevation taxa, which all invaded from low to high altitudes or latitudes (table S1; significant by sign test, $P < 0.001$). Adding range shifts of competitors as a factor improved fit significantly between observed and predicted extinctions ($\Delta\log\text{likelihood} = 45.37$, 1 df, $P < 0.0001$, logistic regression). Therefore, competitive exclusion by invading low-elevation taxa appears to exacerbate climate-change extinctions of high-elevation taxa.

Lizards cannot evolve rapidly enough to track current climate change because of constraints arising from the genetic architecture of thermal preference (12, 13). A phylogenetic correlation between T_b and CT_{max} constrains adaptation. PIC regression of CT_{max} on T_b among Phrynosomatidae suggests that a shift in T_b by 1°C yields only a 0.5°C correlated response in CT_{max} (table S5 and fig. S7). Thus, CT_{max} may not evolve fast enough to keep up with evolved change in T_b . Furthermore, adaptive increase in T_b due to climate change is constrained by genetic correlations in which high T_b necessarily requires prolonged activity out of retreat sites (25), further increasing risk of overheating. Genetic trade-offs with energetically costly traits such as growth (25) also constrain adaptation.

The evolutionary response ($R = h^2s$; s is the selection differential) necessary to keep pace with climate change is further constrained by low heritability for T_b , which we previously estimated at $h^2 = 0.17$ for *Sceloporus occidentalis* in the laboratory (25). We used the physiological model to compute the sustained selection differential at each site j , such that $T_{b,j} + \Delta T_{b,j}$ evolves to match $T_{max,j} + \Delta T_{max,j}$ yielding $\Delta h_{r,j} = 0$ and thereby rescuing population j from extinction [Δ , computed over 1975 to 2009 (historical), 2009 to

2050, and 2050 to 2080]. We assumed $s_j = R_j/h^2 = \Delta T_{b,j}/h^2$, and generation times of 1 year versus 2 years (i.e., lowland versus montane *Sceloporus*, table S1). We expressed these critical levels of adaptive response as surfaces for $s_{sustained}$, the sustained selection differential (Fig. 2B).

We compared the magnitude of selection allowing a species to adapt to climate change with maximum rates sustained under artificial or natural selection (26). Such comparisons are facilitated by dividing each sustained selection differential by the standard deviation ($\sigma_{T_b} = 1.23$ for T_b of Mexican lizards) to obtain i , the standardized intensity of selection (26). Whereas $i > 0.4$ can be sustained in laboratory artificial selection for nine generations (27), studies in nature (26) indicate that $i > 0.4$ computed on an annual basis are rare (<5%). We also reference i to other anthropogenic causes of selection. Overfishing of Atlantic cod yielded $i = 0.55$, among the highest measured, but this selection regime caused demographic collapse of the fishery (28). In México, extinct sites sustained significantly higher i than persistent sites ($i_{extinct} = 0.34 \pm 0.05$ versus $i_{persistent} = 0.13 \pm 0.02$, $t = 4.17$, $P < 0.001$). The relation between intensity of selection and demographic collapse is simple. If sustained for decades, the mortality fraction necessary for selective shifts to new optima compromises population growth rate precipitating local extinction.

If climate change T_{max} continues unabated in México, 56% of viviparous sites will be extinct by 2050 and 66% by 2080 (Fig. 2B). For oviparous sites, 46% will be extinct by 2050 and 61% by 2080. Based on local extinction of all populations surveyed for species, we project 58% species extinction of Mexican *Sceloporus* by 2080. Species extinction (58% by 2080) mirrors local population extinction (61 to 66%) because high-elevation endemics will go completely extinct as widespread lowland taxa expand to high elevations.

We used the model to derive global extinction projections (Fig. 3) for 34 lizard families (Table 1) with 1216 geo-referenced T_b records (table S6). Our data include heliotherms that bask and thermoconformers that do not bask, but track ambient air and surface temperature. T_{max} was obtained from the WorldClim database (29) at 10-arc min resolution (1975, 2020, 2050, and 2080). We used distributional limits of heliothermic lizards of the world in 1975 to calibrate h_r by family, which if exceeded at a given site would precipitate extinction. The extinction model is easily adapted to thermoconformers that maintain T_b close to T_{air} or retreat when $T_{air} > T_{preferred}$. Assuming a sine wave for T_{air} between T_{min} and T_{max} (24-hour period), if the cumulative hours that $T_{air} > T_b$ for a thermoconformer at a given geo-referenced site (table S6) exceeded the h_r of a given lizard family, we assumed it would go extinct. Given $T_{max} - T_b$ at each geo-referenced site, we computed the h_r each species sustained in 1975, and for each family we used

¹Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, CA, 95064, USA. ²Laboratorio de Herpetología, Instituto de Biología, Universidad Nacional Autónoma de México, D.F., 04510, México. ³Department of Biology, Ohio University, 131 Life Sciences Building, Athens, OH 45701, USA. ⁴CNRS UMR 6553, Station Biologique, 35380 Paimpont, France. ⁵Laboratorio de Biología de la Reproducción Animal, Departamento de Biología Comparada, Facultad de Ciencias, Universidad Nacional Autónoma de México, D.F., 04510, México. ⁶Instituto de Ciencias Naturales, Universidad Nacional de Colombia, Sede Bogotá, Colombia. ⁷Instituto de Ecología, A.C., Miguel de Cervantes No. 120 (Cubículo 30C), Complejo Industrial, C.P. 31109, Chihuahua, México. ⁸Centro Nacional Patagónico, Consejo Nacional de Investigaciones Científicas y Técnicas, Blvd. Brown 2915, U9120ACD, Puerto Madryn, Chubut, Argentina. ⁹Museo Nacional de Ciencias Naturales, CSIC, C/ José Gutiérrez, Abascal 2, 28006 Madrid, Spain. ¹⁰Universidad de Concepción, Dpto. Zoología, Casilla 160-C, Concepción, Chile. ¹¹Department of Ecology, Institute of Biology, Universidade do Estado do Rio de Janeiro, Rua São Francisco Xavier 524, Maracanã 20550-019, Rio de Janeiro, Brazil. ¹²Instituto de Investigación en Biodiversidad y Medio Ambiente (INIBIOMA), Consejo Nacional de Investigaciones Científicas y Técnicas, Centro Regional Universitario Bariloche, Universidad Nacional del Comahue, Quintral 1250, San Carlos de Bariloche, Río Negro 8400, Argentina. ¹³Departamento de Herpetología, Museo de Historia Natural, Universidad Nacional Mayor de San Marcos, Av. Arenales 1256, Jesús María Apdo 14-0434, Lima 14, Perú. ¹⁴Laboratoire Ecologie-Evolution, Université UPMC, CNRS UMR 7625, 7 quai Saint Bernard, 75005 Paris, France. ¹⁵Station d'Ecologie Expérimentale du CNRS a Moulis USR 2936, Moulis, 09200 Saint-Girons France. ¹⁶Centre of Excellence in Evolutionary Research, Department of Biological and Environmental Science, Post Office Box 35, FI-40014, University of Jyväskylä, Finland. ¹⁷School of Biological Sciences, Monash University, Victoria 3800, Australia. ¹⁸Department of Biology, Villanova University, 800 Lancaster Avenue, Villanova, PA 19085, USA. ¹⁹Bayworld, Post Office Box i13147, Humewood 6013, South Africa. ²⁰Department of Biology and Bean Life Science Museum, Brigham Young University, Provo, UT 84602, USA.

*To whom correspondence should be addressed. E-mail: lizardrps@gmail.com

†Present address: Laboratoire d'Etude Environnementales des Systèmes Anthropisés (LEESA), UFR Sciences, 2 Bd Lavoisier, 49045 Angers cedex 01, France.

the upper 95% confidence level of h_r (Table 1) as the extinction threshold (iteratively estimated, given global climate surfaces). Calibration with these 1975 distributional limits for *Sceloporus* yields $h_r = 3.9$, which was cross-validated by $h_r = 3.85$ computed from observed extinctions in México (1975 to 2009), and $h_r = 4$, which was estimated directly from T_e at persistent *S. serrifer* sites on the verge of extinction.

Fig. 1. (A) Logistic regression of extinction probability (0 = extant, 1 = extinct) of *Sceloporus* lizards and reproductive mode: $\chi^2 = 7.41$, $P = 0.025$, Δ elevation ($\chi^2 = 8.53$, $P = 0.014$), Δ latitude ($\chi^2 = 7.14$, $P = 0.004$), and Δ longitude (not significant), where Δ refers to deviations from species range midpoints. **(B)** Phylogenetic independent contrasts (PICs) of lineage survival (survival probability of local populations) and Δ elevation ($t = 2.15$, $P = 0.03$), Δ latitude ($t = 3.94$, $P = 0.0001$), and Δ longitude ($t = 2.66$, $P = 0.009$). **(C)** PICs of lineage survival, T_b ($t = 2.32$, $P = 0.02$), T_{air} ($t = 2.31$, $P = 0.02$), and reproductive mode ($t = -2.92$, $P = 0.005$).

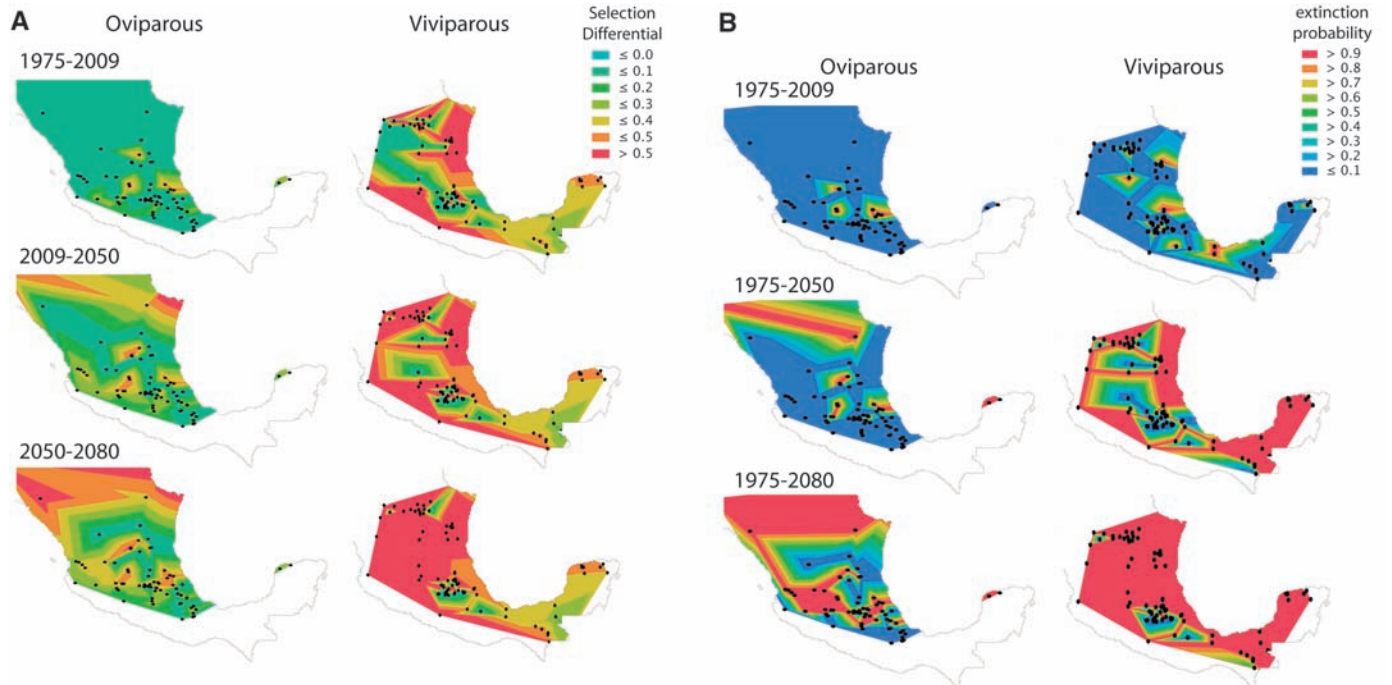
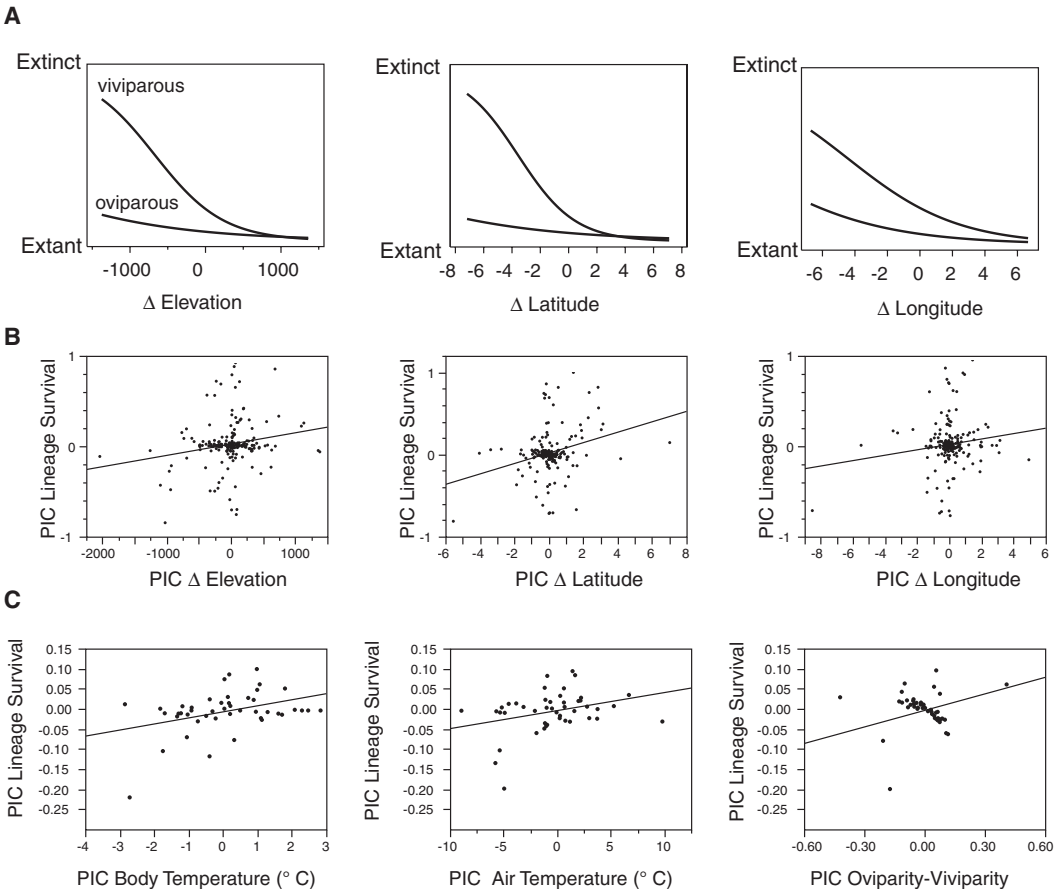


Fig. 2. (A) Sustained selection differentials per year required for T_b to keep pace with global warming. **(B)** Extinctions of Mexican *Sceloporus* lizards (1975 to 2009, 2009 to 2050, 2050 to 2080).

As in the validation of Mexican *Sceloporus* extinction, we computed h_r for temperate lizards over 2 critical reproductive months, but were conservative in modeling critical months required for h_r to be exceeded in the equatorial zone ($\pm 12^\circ$ latitude) where lizards potentially breed year-round (h_r exceeded over 12 months),

and in the wet-dry tropical zone ($\pm 12^\circ$ to 24° latitude: h_r exceeded for 5 to 6 months).

Geo-referenced T_b samples indicate that current (2009) local extinctions average 4% worldwide (Table 1). Global averages will increase fourfold to 16% by 2050 and nearly eightfold to 30% by 2080, while equatorial extinctions will reach 23%

by 2050 and 40% by 2080. Assuming reproduction shifts 1 month earlier in temperate zones [$h^2 = 1.0$ lay date (30)] and proportionately less to the trade zones (i.e., no shift), 2080 global extinctions jump to 38% because spring seasons are warming faster across the globe. Our model is robust to plasticity in T_b (table S7) and initial assump-

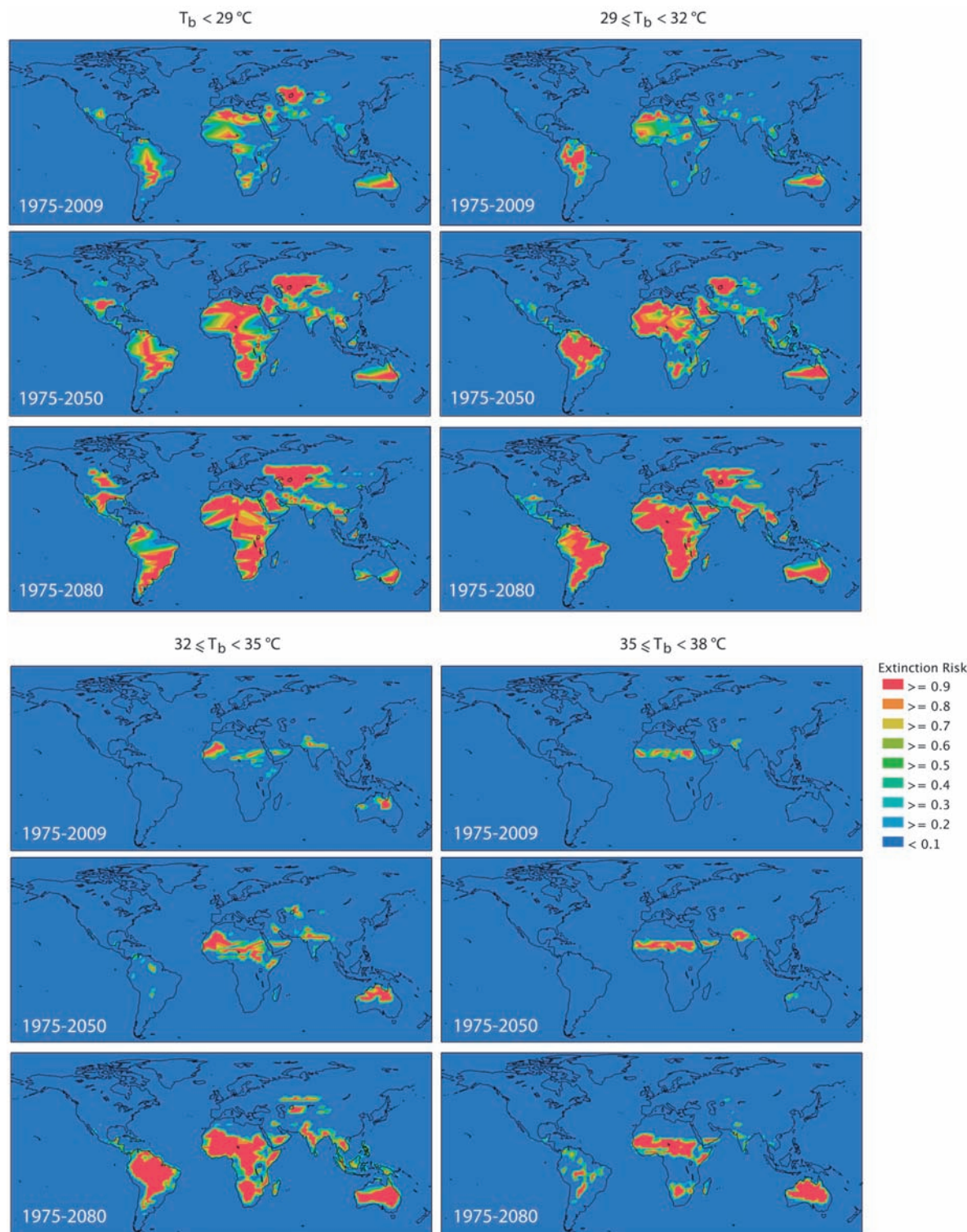


Fig. 3. Contour plots of global levels of local extinction for heliothermic lizards (1975 to 2009, 1975 to 2050, 1975 to 2080), assuming $\bar{h}_r = 4.55$ (23) and various T_b values.

Table 1. Sample size, T_b range, $\bar{T}_b \pm SE$, $\bar{T}_{max}$, h_r , and $n_{species}$ for 34 lizard families. Local extinction rates are based on geo-referenced T_b data and a physiological model of extinction. We also validated model predictions of local extinction risk in 2080 for six families: 57% (± 3 , $n = 200$) for Mexican Phrynosomatidae, 13% (± 2 , $n = 3155$) for South American Liolaemidae, 56% (± 5 , $n = 117$) for European Lacertidae (*L. vivipara*), 13% (± 2 , $n = 1438$) for African Cordylidae + Gerrhosauridae, 57% (± 4 , $n = 125$) on Madagascar, and 10% (± 1 , $n = 2841$) for Australian Egernia Group lizards species. Estimates of species extinctions in each family are derived from the relationships for extinction of all local populations for these six families (table S8).

Family	<i>n</i>	<i>T</i> _b range	$\bar{T}_b \pm SE$	$\bar{T}_{max}$	<i>h</i> _r	Mode of thermoregulation	Local extinction levels				Species extinction	
							<i>n</i> _{spp}	2009	2050	2080	2050	2080
Agamidae	74	19.0–43.8	35.6±0.34	29.8	7.0	Heliothermic	381	0.000	0.169	0.292	0.059	0.266
Amphisbaenidae	2	21.1–21.2	21.2±0.05	28.8	16.2	Fossorial	160	0.000	0.000	0.250	0.000	0.228
Anguidae	10	21.4–32.3	26.7±0.94	20.9	5.6	thermoconformer Heliothermic, a few fossorial	112	0.111	0.111	0.111	0.039	0.101
Annielliidae	2	21.0–23.6	22.3±2.09	20.5	11.5	Fossorial thermoconformer	2	0.000	0.000	0.000	0.000	0.000
Chamaeleonidae	18	22.2–33.5	30.0±0.70	26.8	12.0	Forest thermoconformer	161	0.063	0.063	0.063	0.022	0.057
Cordylidae	11	27.8–33.8	31.5±0.82	23.6	6.8	Heliothermic	54	0.000	0.000	0.200	0.000	0.182
Corytophanidae	4	26.0–35.0	31.9±1.48	29.4	13.4	Forest thermoconformer	9	0.250	0.250	0.250	0.088	0.228
Crotaphytidae	23	35.5–38.9	37.3±0.62	23.3	1.2	Heliothermic	12	0.111	0.167	0.222	0.059	0.202
Carphodactylidae	11	15.1–35.5	24.5±1.59	34.7	10.9	Thermoconformer	30	0.350	0.820	0.820	0.289	0.748
Diplodactylidae	42	16.9–35.9	27.3±0.59	31.2	10.9	Thermoconformer	141	0.070	0.190	0.190	0.067	0.173
Eublepharidae	18	26.6–33.0	28.5±0.44	32.9	10.9	Thermoconformer	28	0.060	0.240	0.240	0.084	0.219
Gekkonidae	40	26.0–35.3	30.1±0.58	32.6	10.9	Thermoconformer	700	0.000	0.000	0.000	0.000	0.000
Phyllodactylidae	13	16.6–38.9	30.6±1.42	30.4	10.9	Thermoconformer	100	0.000	0.000	0.000	0.000	0.000
Pygopodidae	21	24.9–35.1	25.4±0.46	17.9	11.5	Fossorial thermoconformer	38	0.000	0.000	0.000	0.000	0.000
Sphaerodactylidae	19	25.3–38.6	30.2±0.75	33.0	10.9	Thermoconformer	200	0.000	0.000	0.000	0.000	0.000
Gerrhosauridae	4	31.8–33.3	32.6±2.09	28.3	6.8	Heliothermic	16	0.333	0.333	0.333	0.117	0.304
Gymnophthalmidae	20	21.5–29.9	26.4±0.66	30.3	13.8	Leaf litter thermoconformer	193	0.095	0.333	0.667	0.117	0.608
Helodermatidae	2	29.4–30.2	29.8±2.09	24.8	2.7	Heliothermic/thermal inertia	2	0.000	0.000	1.000	0.000	0.912
Iguanidae	20	32.9–42.1	37.3±0.79	28.1	3.7	Heliothermic	36	0.143	0.143	0.286	0.050	0.261
Lacertidae	89	26.7–40.2	35.4±0.31	25.6	3.1	Heliothermic	279	0.034	0.241	0.460	0.085	0.420
Lanthanotidae	1	—	28.0	30.5	9.4	Forest thermoconformer	1	1.000	1.000	1.000	0.352	0.912
Leiocephalidae	1	—	36.3	31.7	2.8	Heliothermic	29	0.000	1.000	1.000	0.352	0.912
Liolaemidae	125	24.4–40.8	33.7±0.27	17.8	1.4	Heliothermic	219	0.027	0.071	0.107	0.025	0.098
Opluridae	3	36.2–39.8	37.7±1.71	31.8	4.0	Heliothermic	7	0.333	0.667	0.667	0.235	0.608
Phrynosomatidae	215	26.8–41.5	35.2±0.20	24.9	3.9	Heliothermic	125	0.037	0.087	0.149	0.031	0.136
Polychrotidae	121	19.6–35.0	29.6±0.27	29.6	14.4	Forest thermoconformer	393	0.018	0.043	0.068	0.015	0.062
Scincidae	210	20.3–38.0	32.9±0.20	26.5	6.2	Heliothermic, a few fossorial	1305	0.015	0.092	0.308	0.032	0.281
Sphenodontidae	1	14.5–21.0	14.5±2.09	18.0	10.7	Nocturnal thermoconformer	1	0.000	0.000	0.000	0.000	0.000
Teiidae	91	26.8–41.3	37.9±0.31	29.0	4.2	Heliothermic	121	0.012	0.136	0.210	0.048	0.192
Trogonophidae	2	22.0–22.5	22.3±0.25	28.8	16.2	Fossorial thermoconformer	8	0.000	0.000	0.000	0.000	0.000
Tropiduridae	72	26.2–38.0	33.7±0.35	28.3	7.7	Heliothermic	111	0.043	0.058	0.087	0.020	0.079
Varanidae	46	28.8–38.9	35.8±0.44	29.7	4.6	Heliothermic/thermal inertia	68	0.001	0.023	0.178	0.008	0.162
Xantusidae	8	18.7–33.0	25.4±1.32	20.7	0.0	Thigmothermic thermoconformer	29	0.000	0.000	0.000	0.000	0.000
Xenosauridae	5	20.3–25.6	23.2±1.48	26.4	11.4	Thigmothermic thermoconformer	6	0.200	0.200	0.600	0.070	0.547

tions made for reproductive periods in the tropics. If h_r for equatorial taxa is computed over the 9 hottest months of reproduction, rather than the conservative assumption of 12 months, global extinctions increase to 39% by 2080.

The global generality of our model is verified by concordant distributions of current observed and predicted local extinctions of lizard biotas from four other continents (table S7). Our model pinpoints exact locations of two Liolaemid species going extinct in South America (*Liolaemus lutzae*, *Phymaturus tenebrosus*: $\chi^2 = 32.1$, $P < 0.0001$). In addition, the model predicts recent (2009) extinctions among 24 resurveyed populations of *L. lutzae* ($\chi^2 = 8.8$, $P = 0.003$). In Europe, our

resurvey of *Lacerta vivipara* revealed 14 extinct sites out of 46 (30%), which are predicted quite precisely by the model ($\chi^2 = 24.4$, $P < 0.001$). In Australia, the model pinpoints 2009 extinctions of *Liopholis slateri* ($\chi^2 = 17.8$, $P < 0.00001$) and 2009 extinctions of *Liopholis kintorei* ($\chi^2 = 3.93$, $P = 0.047$). In Africa, analysis of Gerrhosauridae and Cordylidae at 165 sites predicts <1% extinctions, and yet the model pinpoints the single extinction reported by 2009 (exact P -value = 0.006). We temper this value with extinction projections of 23% for 2009 at Malagasy Gerrhosauridae sites, which is validated by the observed 21% levels of local extinction across several lizard families in Madagascar nature reserves (23).

Thermoconforming lizards have been posited (31) to be more vulnerable to climate change relative to heliotherms. Even though $\bar{T}_b$ of thermoconformers ($27.5^\circ\text{C} \pm 1.8^\circ$) is significantly less than $\bar{T}_b$ of heliotherms ($33.5^\circ\text{C} \pm 1.3$, $t = 2.66$, $P < 0.02$, $n = 34$ families; Table 1), PICs show that extinction risk was unrelated to thermoregulatory mode (fig. S8), but was significantly increased by low $\bar{T}_b$, low h_r , and high $\bar{T}_{\max}$. The similar level of local extinctions in 2009 for Malagasy thermoconformers (21%, $n = 63$) and heliotherms [21%, $n = 34$; (23)] supports this view. Evolved changes in thermoregulatory mode, T_b , h_r , lay date, and habitat preference set risk as $T_{\max}$ rises, but owing to trade-offs, T_b and h_r cannot be simultaneously maximized, hence extinction risk is independent of mode (fig. S8). Moreover, extinction risk is not higher for conformers because heliotherms inhabit equatorial regions (i.e., sub-Saharan Africa) that are unavailable to thermoconformers [a factor not considered by (31) or other models (10)], and these areas are warming rapidly (Fig. 3).

Our model, based on T_b , h_r in activity during reproduction, and timing of breeding, assesses salient adaptations that affect thermal extinctions. Concordant verification of 2009 levels of local lizard extinction in North and South America, Europe, Africa, and Australia confirm that extinctions span tropical, temperate, rainforest, and desert habitats. Estimates of evolutionary rates required to keep pace with global change indicate that sustained and intense selection compromises population growth rates, precipitating extinctions. Probability of local extinction is projected to result in species extinction probabilities of 6% by 2050 and 20% by 2080 (table S8). Range shifts only trivially offset losses, because widespread species with high T_b shift to ranges of endemics, thereby accelerating their demise. Although global efforts to reduce CO_2 may avert 2080 scenarios, 2050 projections are unlikely to be avoided; deceleration in $\bar{T}_{\max}$ lags atmospheric CO_2 storage by decades (4). Therefore, our findings indicate that lizards have already crossed a threshold for extinctions.

References and Notes

- C. D. Thomas *et al.*, *Nature* **427**, 145 (2004).
- J. A. Pounds, R. Puschendorf, *Nature* **427**, 107 (2004).
- R. J. Wilson *et al.*, *Ecol. Lett.* **8**, 1138 (2005).
- B. Hare, M. Meinshausen, *Clim. Change* **75**, 111 (2006).
- D. A. Stainforth *et al.*, *Nature* **433**, 403 (2005).
- P. A. Stott, J. A. Kettleborough, *Nature* **416**, 723 (2002).
- C. A. Deutsch *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 6668 (2008).
- J. J. Tewksbury, R. B. Huey, C. A. Deutsch, *Science* **320**, 1296 (2008).
- M. B. Araújo, R. J. Whittaker, R. J. Ladle, M. Erhard, *Glob. Ecol. Biogeogr.* **14**, 529 (2005).
- J. Harte, A. Ostling, J. L. Green, A. Kinzig, *Nature* **430**, 3, 33, discussion 33 (2004).
- M. Kearney, R. Shine, W. P. Porter, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 3835 (2009).
- R. B. Huey, P. E. Hertz, B. Sinervo, *Am. Nat.* **161**, 357 (2003).
- J. R. Etterson, R. G. Shaw, *Science* **294**, 151 (2001).
- R. B. Huey, R. D. Stevenson, *Am. Zool.* **19**, 357 (1979).
- W. P. Porter, *Physiol. Zool.* **62**, 286 (1989).
- L. J. Guillelte Jr., *Bioscience* **43**, 742 (1993).
- C. A. Beuchat, *Copeia* **1986**, 971 (1986).
- W. P. Maddison, D. R. Maddison, www.mesquiteproject.org (2008).
- P. E. Midford, T. Garland Jr., W. P. Maddison. (2005).
- J. H. Brown, *Am. Nat.* **124**, 255 (1984).
- J. Terborgh, *Am. Nat.* **107**, 481 (1973).
- E. M. Dzialowski, *J. Therm. Biol.* **30**, 317 (2005).
- Materials and Methods are available as supporting material on Science Online.
- M. Massot, J. Clobert, R. Ferrière, *Glob. Change Biol.* **14**, 461 (2008).
- B. Sinervo, *Oecologia* **83**, 228 (1990).
- H. E. Hoekstra *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 9157 (2001).
- P. J. Berger, W. R. Harvey, *J. Anim. Sci.* **40**, 38 (1975).
- D. P. Swain, A. F. Sinclair, J. M. Hanson, *Proc. Biol. Sci.* **274**, 1015 (2007).
- R. J. Hijmans, S. E. Cameron, J. L. Parra, P. G. Jones, A. Jarvis, *Int. J. Clim.* **25**, 1965 (2005).
- B. Sinervo, P. Doughty, *Evolution* **50**, 1314 (1996).
- R. B. Huey *et al.*, *Proc. Biol. Sci.* **276**, 1939 (2009).
- Research of B.S. was funded by the National Geographic Society, UC Mexus, UCSC Committee-On-Research, NSF awards (DEB 0108577, IBN 0213179, LTREB DEB 051597), CNRS fellowships, and visiting professorships (Museum Nationale d'Histoire Naturelle, Université Paris 6, Université Paul Sabatier Toulouse III), PAPIIT-UNAM IN213405 and 224208 to F.M.-C., a Université Paul Sabatier Toulouse III Visiting Professorship to D.B.M., CONACYT grants (4171N and 52852Q) to M.V.-S.C., grant CONACYT-SEP (43142-Q) to H.G., a CONACYT fellowship to R.N.M.-L., CNRS funding to B.H., and M.M., Biodivera: Tenlamas and from ANR Blanche: DIAME to J.C., CONICET grants to L.J.A. and M.M., FONDYCET 1090664 grants to P.V.S., CGL2005-03156 and CLG2008-04164 grant from SMSI to I.J.R., APCT-PIC1086 grant to N.I., scholarships and grants from Universidad Nacional Autónoma de México and American Museum of Natural History to M.V.-S.C., Academy of Finland grant (108955) to T.A.O., Australian Research Council grants to D.G.C., NSF awards DEB 0515909 and 0844523 to A.M.B., NSF award OISE 0530267, PIRE-Patagonia grant to J.W.S., L.J.A., M.M., and P.V.S. and Brigham Young University funding (Biology Department, Kennedy Center for International Studies, Bean Life Science Museum) to J.W.S.

Supporting Online Material

www.sciencemag.org/cgi/content/full/328/5980/894/DC1
Materials and Methods
Figs. S1 to S9
Tables S1 to S8
References

16 November 2009; accepted 7 April 2010
10.1126/science.1184695

Carbon Dioxide Enrichment Inhibits Nitrate Assimilation in Wheat and *Arabidopsis*

Arnold J. Bloom,* Martin Burger,† Jose Salvador Rubio Asensio, Asaph B. Cousins‡

The concentration of carbon dioxide in Earth's atmosphere may double by the end of the 21st century. The response of higher plants to a carbon dioxide doubling often includes a decline in their nitrogen status, but the reasons for this decline have been uncertain. We used five independent methods with wheat and *Arabidopsis* to show that atmospheric carbon dioxide enrichment inhibited the assimilation of nitrate into organic nitrogen compounds. This inhibition may be largely responsible for carbon dioxide acclimation, the decrease in photosynthesis and growth of plants conducting C_3 carbon fixation after long exposures (days to years) to carbon dioxide enrichment. These results suggest that the relative availability of soil ammonium and nitrate to most plants will become increasingly important in determining their productivity as well as their quality as food.

The concentration of CO_2 in Earth's atmosphere has increased from about 280 to 390 $\mu\text{mol CO}_2$ per mol of atmosphere ($\mu\text{mol mol}^{-1}$) since 1800, and predictions are that it will reach between 530 and 970 $\mu\text{mol mol}^{-1}$ by the end of the 21st century (1). Plants could mitigate these changes through photosynthetic conversion of atmospheric CO_2 into carbohydrates and other organic compounds, yet the potential for this mitigation remains uncertain. Photorespiration is the biochemical pathway in which the chloroplast enzyme Rubisco catalyzes the oxidation of the

high-energy substrate RuBP rather than catalyzes the carboxylation of RuBP through the C_3 carbon-fixation pathway (2). Elevated CO_2 (or

Department of Plant Sciences, University of California at Davis, Davis, CA 95616, USA.

*To whom correspondence should be addressed. E-mail: ajbloom@ucdavis.edu

†Present address: Department of Land, Air and Water Resources, University of California at Davis, Davis, CA 95616, USA.

‡Present address: School of Biological Sciences, Post Office Box 646340, Washington State University, Pullman, WA 99164-6340, USA.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

The following resources related to this article are available online at www.sciencemag.org (this information is current as of May 13, 2010):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:
<http://www.sciencemag.org/cgi/content/full/328/5980/899>

Supporting Online Material can be found at:
<http://www.sciencemag.org/cgi/content/full/328/5980/899/DC1>

This article **cites 26 articles**, 7 of which can be accessed for free:
<http://www.sciencemag.org/cgi/content/full/328/5980/899#otherarticles>

This article appears in the following **subject collections**:
Ecology
<http://www.sciencemag.org/cgi/collection/ecology>

low O_2) atmospheric concentrations decrease rates of photorespiration and initially enhance rates of photosynthesis and growth by as much as 35% in most plants (C_3 plants). This enhancement, however, diminishes over time (days to years), a phenomenon known as CO_2 acclimation (3, 4). Most studies suggest a strong link between CO_2 acclimation and plant nitrogen status [for example, (5)].

Nitrogen is the mineral element that organisms require in greatest quantity (6). The primary source of N for terrestrial plants is soil inorganic N in the forms of nitrate (NO_3^-) and ammonium (NH_4^+). Root absorption of NO_3^- and NH_4^+ from the soil and assimilation of NO_3^- and NH_4^+ into organic N compounds within plant tissues have a large influence on primary productivity. The assimilation of NO_3^- involves the sequential conversion of NO_3^- into NO_2^- , then into NH_4^+ , then into glutamine, and finally into other organic N compounds. The first step of this process occurs in the cytosol, and the subsequent ones occur within chloroplasts or plastids.

Previously, we reported that atmospheric CO_2 enrichment does not stimulate the growth of wheat plants receiving NO_3^- as a sole N source to the same extent as those receiving NH_4^+ (7). This result, as well as gas exchange measurements of wheat and *Arabidopsis*, suggested that elevated CO_2 (or low O_2) atmospheric concentrations, which are conditions that decrease photorespiration, inhibit NO_3^- assimilation in the shoots of C_3 plants (7, 8). In the present study, we assessed the influence of

elevated CO_2 and sometimes low O_2 atmospheric concentrations on NO_3^- assimilation in wheat and *Arabidopsis* using several independent methods.

Our first method, NO_3^- depletion, involved growing plants under ambient CO_2 and O_2 conditions and depriving them of NO_3^- nutrition until their tissue NO_3^- contents decreased to low steady levels (9). The shoots of these NO_3^- -depleted plants were then subjected to an ambient or elevated CO_2 atmospheric concentration and an ambient or low O_2 atmospheric concentration. The roots then received a pulse of NO_3^- in the nutrient medium. The decline of NO_3^- concentrations in the medium provided an estimate of net plant NO_3^- absorption, and the difference between this net NO_3^- absorption and the accumulation of free NO_3^- within the plants provided an estimate of plant NO_3^- assimilation (8).

By the NO_3^- -depletion method, wheat assimilated nearly all of the NO_3^- that its roots absorbed, whereas *Arabidopsis* assimilated less than half of the NO_3^- that its roots absorbed (Fig. 1). Estimates of NO_3^- absorption via this method were slower than estimates via the isotopic methods (^{15}N and ^{14}N labeling) that are described below. The NO_3^- -depletion method deprives plants of NO_3^- for several days, and this has been shown to down-regulate the expression of several NO_3^- transporters (10). In contrast, the isotopic methods maintain a constant NO_3^- concentration in the medium and would not alter the expression of transporters.

The NO_3^- -depletion method showed that an elevated CO_2 atmospheric concentration around the shoots decreased the rate of NO_3^- assimila-

tion (Fig. 1). A low O_2 atmospheric concentration also decreased the rate of NO_3^- assimilation, but the decrease in *Arabidopsis* was not statistically significant. By this method, as in the isotopic methods described below, NO_3^- absorption varied with elevated CO_2 and low O_2 in a pattern that was similar to NO_3^- assimilation.

Our second method for assessing NO_3^- assimilation, ^{15}N labeling, entailed growing plants under ambient CO_2 and O_2 conditions in a hydroponic medium containing 0.2 mM NO_3^- at natural abundance levels of N isotopes ($\approx 0.366\%$ ^{15}N). We then shifted the plants to an ambient or elevated CO_2 atmospheric concentration and an ambient or low O_2 atmospheric concentration and to a root medium containing 0.2 mM NO_3^- that was 25%-enriched in ^{15}N - NO_3^- . After a 12-hour labeling period, we analyzed the plant tissues for ^{15}N enrichment of total N and free NO_3^- ; the ^{15}N enrichment of total N provided an estimate of net ^{15}N absorption, and the difference between the ^{15}N enrichment of total N and that of free NO_3^- provided an estimate of $^{15}NO_3^-$ assimilation.

According to ^{15}N labeling, wheat and *Arabidopsis* assimilated about two-thirds of the ^{15}N - NO_3^- they absorbed (Fig. 1). In wheat, net $^{15}NO_3^-$ absorption and assimilation were significantly greater under ambient CO_2 and O_2 atmospheric concentrations than under an elevated CO_2 or low O_2 concentration. In *Arabidopsis*, net $^{15}NO_3^-$ assimilation was significantly greater under an ambient CO_2 and O_2 atmospheric concentration than under an elevated CO_2 concentration.

In our third method, ^{14}N - NO_3^- labeling, we grew plants under ambient CO_2 and O_2 conditions in a hydroponic medium that contained 99.9% enriched ^{15}N - NO_3^- as the sole N source. When the wheat and *Arabidopsis* plants were about 14 and 36 days old, respectively, the shoots were exposed to an ambient or elevated CO_2 atmospheric concentration and an ambient or low O_2 atmospheric

Fig. 1. Three methods for assessing nitrate absorption (Absorb) and assimilation (Assim.) in wheat and *Arabidopsis* plants where the shoots were exposed to atmospheres containing 380 $\mu\text{mol mol}^{-1}$ CO_2 and 21% O_2 , 720 $\mu\text{mol mol}^{-1}$ CO_2 and 21% O_2 , or 380 $\mu\text{mol mol}^{-1}$ CO_2 and 2% O_2 . Shown are means $\pm$ SE ($n = 6$ to 18) in $\mu\text{mol NO}_3^-$ per gram plant per minute. Within a species and for absorption separately from assimilation, bars labeled with different letters differ significantly ($P < 0.05$). Data for the NO_3^- -depletion method include those from an earlier study (8).

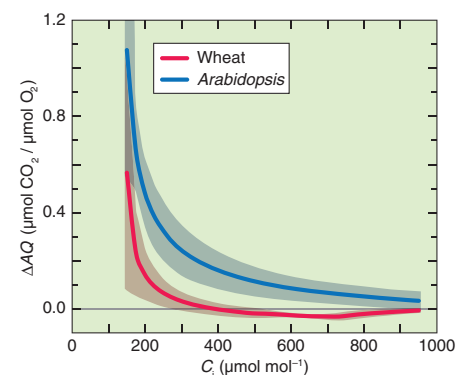
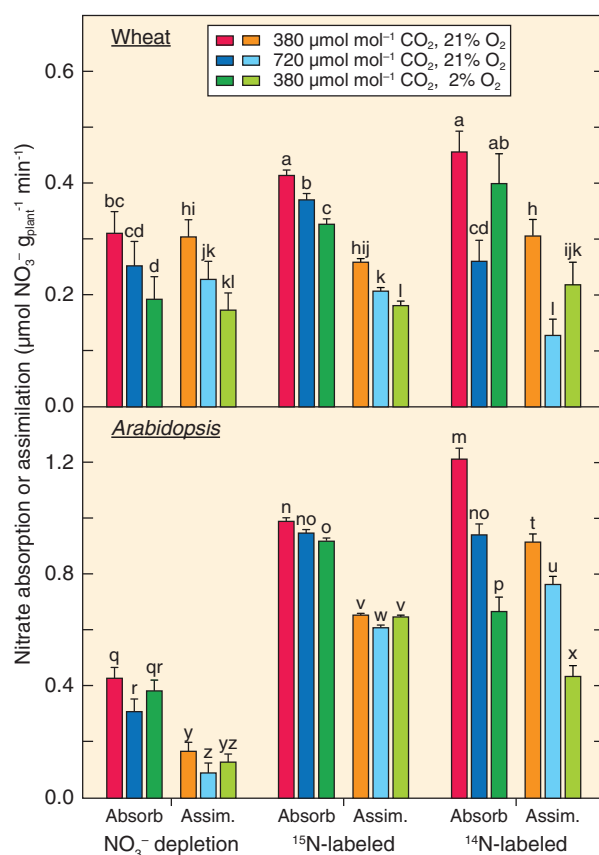


Fig. 2. The ΔAQ , the change in the ratio of shoot CO_2 consumption to O_2 evolution with a shift from NO_3^- to NH_4^+ nutrition, as a function of shoot internal CO_2 concentration (C_i) in wheat and *Arabidopsis*. An instrumental system described previously (12) monitored shoot gas fluxes. A biochemical model of photosynthesis (33, 34), which we fitted to the data, interpolated the values at regular C_i intervals. Shown are the means $\pm$ SE ($n = 8$ for wheat and 4 for *Arabidopsis*).

concentration. After a few hours under these atmospheric conditions, the roots received a pulse of NO_3^- containing the isotopes at their natural abundance levels of nitrogen (99.633% ^{14}N). We estimated ^{14}N - NO_3^- absorption and assimilation from the decreases in the ^{15}N enrichment of total N and free NO_3^- in plant tissues after a 12-hour exposure to ^{14}N - NO_3^- .

Differences in atmospheric CO_2 or O_2 concentration produced distinct patterns of the ^{14}N labeling (Fig. 1). Wheat and *Arabidopsis* assimilated about two-thirds of the ^{14}N - NO_3^- they absorbed. In both species, an elevated CO_2 or low O_2 atmospheric concentration significantly decreased ^{14}N - NO_3^- assimilation.

Our fourth method for assessing NO_3^- assimilation depended on the assimilatory quotient (AQ), the ratio of net CO_2 consumption to net O_2 evolution from shoots during photosynthesis. Values of AQ decrease as NO_3^- assimilation increases, because additional electrons generated from the light-dependent reactions of photosynthesis are transferred first to NO_3^- and then to NO_2^- . This stimulates net O_2 evolution but has

little effect on CO_2 consumption (7, 8, 11–13). The ΔAQ —the difference in the AQ between plants receiving NH_4^+ as the sole N source and those receiving NO_3^- —is strongly correlated with shoot NO_3^- assimilation. For example, ΔAQ did not deviate significantly from zero in plants with relatively low NO_3^- reductase activities (that is, *Arabidopsis* knockout mutants and 48-day-old wild-type *Arabidopsis*), whereas ΔAQ was positive in plants with significant NO_3^- reductase activities (that is, wild-type wheat, transgenic *Arabidopsis* that overexpresses NO_3^- reductase, and 36-day-old wild-type *Arabidopsis*) (8). Rates of photorespiration do not influence AQ or ΔAQ , because this process changes neither net CO_2 consumption nor net O_2 evolution (2).

During these gas-exchange measurements, we exposed the shoots of wheat and *Arabidopsis* to various atmospheric CO_2 concentrations. To account for differences in stomatal conductance, we expressed shoot CO_2 and O_2 fluxes as a function of apparent shoot internal CO_2 concentrations (C_i) that we calculated from water vapor exchange. With increasing C_i , the ΔAQ in both

species declined from a positive value to one that was not significantly different from zero (Fig. 2), indicating that NO_3^- assimilation was significant at subambient and ambient CO_2 concentrations but was negligible at elevated CO_2 concentrations.

Our fifth method relied on the isotopic discrimination of NO_3^- reductase. When both isotopic forms of NO_3^- are readily available, this enzyme preferentially converts ^{14}N - NO_3^- into organic N compounds by about 15 per mil (‰) (14), and so organic N compounds in plant tissues are more depleted in ^{15}N than the N compounds in the growth medium are. In contrast, when NO_3^- availability limits assimilation, this enzyme discriminates less against ^{15}N - NO_3^- and assimilates relatively more ^{15}N - NO_3^- , and so organic N compounds in plant tissues become more enriched in ^{15}N .

We grew wheat and *Arabidopsis* for 14 and 22 days, respectively, in a medium that contained 0.2 or 1.0 mM NO_3^- ($\delta^{15}\text{N} = -4\text{‰}$) as the sole N source and under an ambient or elevated CO_2 atmospheric concentration. The difference between total plant N and free NO_3^- in the tissues provided an estimate of organic N. Wheat shoot growth did not respond to any of the treatments, whereas *Arabidopsis* shoot growth was greater at the higher NO_3^- level but did not respond to CO_2 treatment (Fig. 3). Shoot NO_3^- contents were higher in plants that were grown at the higher NO_3^- level, and shoot organic N contents decreased under CO_2 enrichment, although the decrease in wheat grown at 1.0 mM was not statistically significant (Fig. 3).

Shoot $\delta^{15}\text{N}$ of organic N (Fig. 3) and plant NO_3^- assimilation rate assessed via the isotopic methods (Fig. 1) were lower in wheat than in *Arabidopsis*. In both species, $\delta^{15}\text{N}$ of shoot organic N decreased in plants grown at the higher NO_3^- level or under CO_2 enrichment, although the decrease with CO_2 in *Arabidopsis* that was grown at 1.0 mM was not statistically significant (Fig. 3). All of these results are consistent with NO_3^- assimilation discriminating more against ^{15}N - NO_3^- when the availability of NO_3^- at the sites of NO_3^- reduction was higher, as a consequence of either slower NO_3^- assimilation in a species (wheat versus *Arabidopsis*), increased NO_3^- supply (1.0 versus 0.2 mM), or decreased NO_3^- assimilation (elevated versus ambient CO_2).

In this study, five independent methods affirm that CO_2 enrichment inhibits NO_3^- assimilation in wheat and *Arabidopsis* plants. The predominant form of N available to plants in most environments is NO_3^- (6); therefore, CO_2 inhibition of NO_3^- assimilation would lead to lower organic N production. Indeed, this could be responsible for the 7.4 to 11% decrease in wheat grain protein (15, 16) and the 20% decrease in total protein content of *A. thaliana* (Columbia) (17) observed under CO_2 enrichment in FACE (free-air CO_2 enrichment) experiments. Because the influence of CO_2 enrichment on leaf organic N contents is highly correlated with its influence on photosynthesis and growth (5), it is reasonable to assume that CO_2 inhibition of NO_3^- assimila-

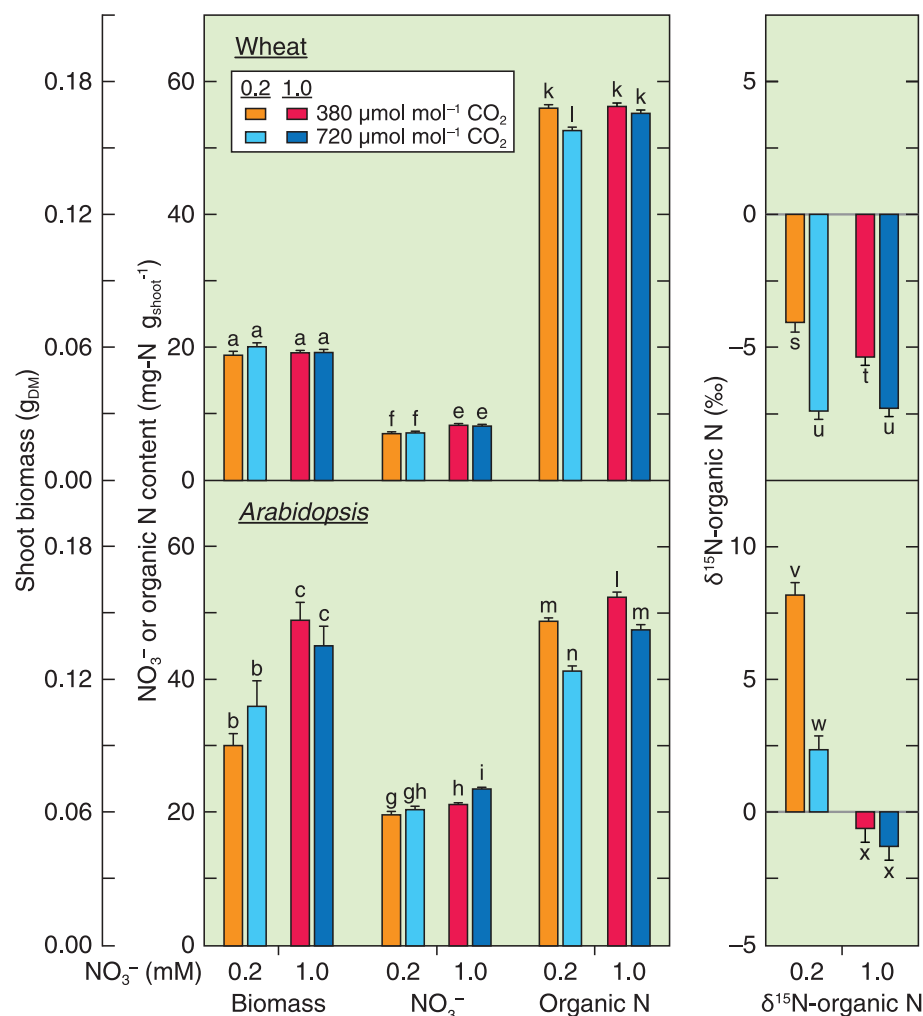


Fig. 3. Shoot biomass, NO_3^- content, organic N content, and $\delta^{15}\text{N}$ of organic N in wheat (upper panels) and *Arabidopsis* (lower panels) grown at 0.2 or 1.0 mM NO_3^- and 380 or 720 $\mu\text{mol mol}^{-1}$ CO_2 . Shown are means $\pm$ SE ($n = 6$ to 12). For each parameter, bars labeled with different letters differ significantly ($P < 0.05$).

tion and the resultant decline in plant organic N contents play a major role in the phenomenon of CO₂ acclimation, the decline of photosynthesis, and growth of C₃ plants after long exposures (days to years) to CO₂ enrichment.

The extent to which plants use NO₃[−] versus NH₄⁺ as N sources varies over seasons, years, locations, and species (6). This variation in the relative dependence on NO₃[−] could explain the observed variation in CO₂ acclimation. Net primary productivity diminished under CO₂ enrichment in an annual California grassland for which NO₃[−] was the predominant N source (18), presumably because NO₃[−] assimilation was inhibited and plant organic N compounds became limiting. In contrast, *Scirpus olneyi*, the prominent C₃ plant in the Chesapeake Bay marsh, which is an NH₄⁺-dominated ecosystem, showed little CO₂ acclimation. Even after a decade of treatment, photosynthesis and growth of this species remained about 35% greater under CO₂ enrichment (19), with little change in N contents (20).

Root NO₃[−] absorption and plant NO₃[−] assimilation were generally correlated in the N-depletion and isotopic methods, yet differences in the responses to elevated CO₂ or low O₂ atmospheric concentrations were sometimes significant for assimilation but not for absorption (Fig. 1). Moreover, CO₂ enrichment decreased shoot organic N contents, but did not change or even increased shoot NO₃[−] contents (Fig. 3). CO₂ enrichment also increased ¹⁵N isotope discrimination during NO₃[−] assimilation (Fig. 3), indicating that NO₃[−] availability became less limiting to assimilation. Finally, changes in atmospheric CO₂ concentration influenced shoot NO₃[−] assimilation within minutes (time response of Fig. 2 not shown). These CO₂ changes also influenced transpiration rapidly, but root NO₃[−] and NH₄⁺ absorption from well-mixed hydroponic solutions is independent of transpiration (21). Together, these results indicate that elevated CO₂ or low O₂ atmospheric concentrations inhibited NO₃[−] assimilation and that assimilation controlled root absorption, rather than elevated CO₂ or low O₂ atmospheric concentrations influencing root NO₃[−] absorption directly.

One physiological mechanism that may be responsible for the relationship between elevated CO₂ or low O₂ atmospheric concentrations and NO₃[−] assimilation involves the first biochemical step of NO₃[−] assimilation, the conversion of NO₃[−] to NO₂[−] in the cytoplasm of leaf mesophyll cells. Photorespiration stimulates the export of malic acid from chloroplasts (22) and increases the availability of the reduced form of nicotinamide adenine dinucleotide (NADH) in the cytoplasm (23) that powers this first step (24, 25). Elevated CO₂ or low O₂ atmospheric concentrations decrease photorespiration and thereby decrease the amount of reductant available to power NO₃[−] reduction. In contrast, the C₄ carbon fixation pathway generates ample amounts of malic acid and NADH in the cytoplasm of mesophyll cells. This may explain why shoot NO₃[−] assimilation is relatively independent of

CO₂ concentrations in C₄ plants (26) and limited to the mesophyll (27).

Another physiological mechanism that may link NO₃[−] assimilation and elevated CO₂ is NO₂[−] translocation from the cytosol into the chloroplast. Six transporters of the *Nar1* family are involved in NO₂[−] translocation from the cytosol into the chloroplast in *Chlamydomonas*, and some of these transport both NO₂[−] and HCO₃[−] (28). We have shown that HCO₃[−] inhibits NO₂[−] influx into isolated wheat and pea chloroplasts (7), indicating that an analogous system is operating in higher plants. Slower NO₂[−] influx into the chloroplast under CO₂ enrichment would decrease NO₃[−] assimilation.

A third physiological mechanism linking CO₂ enrichment and NO₃[−] assimilation involves competition for reductant in the chloroplast stroma. Several processes within the stroma—C₃ carbon fixation, the reduction of NO₂[−] to NH₄⁺, and the incorporation of NH₄⁺ into amino acids—require reduced ferredoxin generated by photosynthetic electron transport. Key enzymes in these processes have different affinities for reduced ferredoxin: FNR (ferredoxin-NADP reductase) has a Michaelis constant *K_m* of 0.1 μM, NiR (nitrite reductase) has a *K_m* of 0.6 μM, and GOGAT (glutamate synthase) has a *K_m* of 60 μM (29). As a result, NO₃[−] assimilation may proceed only if the availability of reduced ferredoxin exceeds that needed for the formation of the reduced form of nicotinamide adenine dinucleotide phosphate (NADPH) (24, 30). For most plants, this occurs when CO₂ availability limits C₃ carbon fixation (7).

The phenomenon of CO₂ acclimation may have several explanations. According to the carbohydrate sink limitation hypothesis, plants under CO₂ enrichment initially assimilate more CO₂ into carbohydrates than they can incorporate into their growing tissues; in response, they diminish CO₂ assimilation by decreasing their levels of Rubisco and other proteins (3). An alternative explanation is the progressive N limitation hypothesis in which shoots accumulate carbohydrates faster than plants can acquire N, making leaf N contents decrease (4, 31, 32). As these leaves senesce and drop to the ground, plant litter quality declines, microbial immobilization of soil N increases because of the high C-to-N ratios in the litter, soil N availability to plants further diminishes because more soil N is tied up in microorganisms, plants become even more N limited, plant protein levels decline, and plant processes including photosynthesis and growth slow down.

Both of these hypotheses about CO₂ acclimation fit nicely into the framework of our results. The decline in Rubisco predicted by the carbohydrate sink hypothesis might derive from CO₂ inhibition of NO₃[−] assimilation and the subsequent decline in plant organic N compounds (Fig. 3). The decline in leaf N contents predicted by the progressive N limitation hypothesis might derive from CO₂ inhibition of NO₃[−] assimilation and the subsequent decline in plant NO₃[−] absorption (Fig. 1).

Our findings have implications for food production. Nitrate is the most abundant form of N in

agricultural soils (6). As atmospheric CO₂ concentrations rise and NO₃[−] assimilation diminishes, crops will become depleted of organic N compounds (see Fig. 3), including protein, and food quality will suffer. Increasing nitrogen fertilization might compensate for slower NO₃[−] assimilation rates (Fig. 3), but such fertilization rates might not be economically or environmentally feasible. Greater reliance on NH₄⁺ fertilizers and inhibitors of nitrification (microbial conversion of NH₄⁺ to NO₃[−]) might avoid the bottleneck of NO₃[−] assimilation, but would require sophisticated fertilizer management to prevent NH₄⁺ toxicity, which occurs when free NH₄⁺ accumulates in plant tissues if they absorb more of this compound than they can assimilate into amino acids. To address these issues, a better understanding of plant NH₄⁺ and NO₃[−] assimilation is critical.

References and Notes

- IPCC, in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, S. Solomon et al., Eds. (Cambridge Univ. Press, Cambridge, 2007), pp. 1–18.
- C. H. Foyer, A. J. Bloom, G. Queval, G. Noctor, *Annu. Rev. Plant Biol.* **60**, 455 (2009).
- S. P. Long, E. A. Ainsworth, A. Rogers, D. R. Ort, *Annu. Rev. Plant Biol.* **55**, 591 (2004).
- P. B. Reich et al., *Nature* **440**, 922 (2006).
- D. S. Ellsworth et al., *Glob. Change Biol.* **10**, 2121 (2004).
- E. Epstein, A. J. Bloom, *Mineral Nutrition of Plants: Principles and Perspectives* (Sinauer Associates, Sunderland, MA, ed. 2, 2005).
- A. J. Bloom, D. R. Smart, D. T. Nguyen, P. S. Searles, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 1730 (2002).
- S. Rachmilevitch, A. B. Cousins, A. J. Bloom, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 11506 (2004).
- Materials and methods are available as supporting material on Science Online.
- N. M. Crawford, B. J. Forde, in *The Arabidopsis Book*, C. Somerville, E. Meyerowitz, Eds. (American Society of Plant Physiologists, Rockville, MD, 2002); www.aspb.org/publications/arabidopsis/.
- J. Myers, in *Photosynthesis in Plants*, J. Franck, W. E. Loomis, Eds. (Iowa State College Press, Ames, IA, 1949), pp. 349–364.
- A. J. Bloom, R. M. Caldwell, J. Finazzo, R. L. Warner, J. Weissbart, *Plant Physiol.* **91**, 352 (1989).
- Y.-P. Cen, D. H. Turpin, D. B. Layzell, *Plant Physiol.* **126**, 1555 (2001).
- G. Tcherkez, G. D. Farquhar, *Funct. Plant Biol.* **33**, 531 (2006).
- B. A. Kimball et al., *New Phytol.* **150**, 295 (2001).
- P. Högy et al., *Plant Biol.* **11** (suppl. 1), 60 (2009).
- P. H. Li et al., *Plant Mol. Biol.* **62**, 593 (2006).
- J. S. Dukes et al., *PLoS Biol.* **3**, 1829 (2005).
- D. P. Rasse, G. Peresta, B. G. Drake, *Glob. Change Biol.* **11**, 369 (2005).
- J. E. Erickson, J. P. Magonigal, G. Peresta, B. G. Drake, *Glob. Change Biol.* **13**, 202 (2007).
- E. D. Schulze, A. J. Bloom, *Plant Physiol.* **76**, 827 (1984).
- J. E. Backhausen et al., *Planta* **207**, 105 (1998).
- A. U. Igamberdiev, N. V. Bykova, P. J. Lea, P. Gardeström, *Physiol. Plant.* **111**, 427 (2001).
- J. M. Robinson, in *Models in Plant Physiology and Biochemistry*, D. W. Newman, K. G. Stuart, Eds. (CRC Press, Boca Raton, FL, 1987), vol. 1, pp. 25–35.
- A. Quesada, E. la Gómez-García, E. Fernández, *Trends Plant Sci.* **5**, 463 (2000).
- A. B. Cousins, A. J. Bloom, *Plant Cell Environ.* **26**, 1525 (2003).
- C. K. M. Rathnam, G. E. Edwards, *Plant Physiol.* **57**, 881 (1976).

28. V. Mariscal *et al.*, *Protist* **157**, 421 (2006).
29. D. B. Knaff, in *Oxygenic Photosynthesis: The Light Reactions*, D. R. Ort, C. F. Yocum, Eds. (Kluwer Academic, Dordrecht, Netherlands, 1996), vol. 4, pp. 333–361.
30. J. E. Backhausen, C. Kitzmann, P. Horton, R. Scheibe, *Photosynth. Res.* **64**, 1 (2000).
31. R. J. Norby, M. F. Cotrufo, P. Ineson, E. G. O'Neill, J. G. Canadell, *Oecologia* **127**, 153 (2001).
32. B. A. Hungate, J. S. Dukes, M. R. Shaw, Y. Q. Luo, C. B. Field, *Science* **302**, 1512 (2003).
33. S. von Caemmerer, *Biochemical Models of Leaf Photosynthesis* (CSIRO Publishing, Collingwood, Australia, 2000).
34. T. D. Sharkey, C. J. Bernacchi, G. D. Farquhar, E. L. Singaas, *Plant Cell Environ.* **30**, 1035 (2007).
35. We thank C. van Kessel, L. Jackson, and E. Carlisle for their review of the manuscript. This research was supported by NSF grants IBN-03-43127 and IOS-08-18435, by the National Research Initiative Competitive grant number 2008-35100-04459 from the U.S. Department of Agriculture National Institute of

Food and Agriculture, and by a postdoctoral fellowship from Agencia Regional de Ciencia y Tecnología, Region de Murcia, Spain to J.S.R.A. A.J.B. is a paid consultant to the Monsanto Corporation.

Supporting Online Material

www.sciencemag.org/cgi/content/full/328/5980/899/DC1
Materials and Methods
References

28 December 2009; accepted 24 March 2010
10.1126/science.1186440

Resource Management Cycles and the Sustainability of Harvested Wildlife Populations

John M. Fryxell,^{1*} Craig Packer,² Kevin McCann,¹ Erling J. Solberg,³ Bernt-Erik Sæther⁴

Constant harvest policies for fish and wildlife populations can lead to population collapse in the face of stochastic variation in population growth rates. Here, we show that weak compensatory response by resource users or managers to changing levels of resource abundance can readily induce harvest cycles that accentuate the risk of catastrophic population collapse. Dynamic system models incorporating this mix of feedback predict that cycles or quasi-cycles with decadal periodicity should commonly occur in harvested wildlife populations, with effort and quotas lagging far behind resources, whereas harvests should exhibit lags of intermediate length. Empirical data gathered from three hunted populations of white-tailed deer and moose were consistent with these predictions of both underlying behavioral causes and dynamical consequences.

One of the most central problems in ecology is what causes some harvested populations to collapse, whereas others are able to withstand exploitation (1–4)? Population collapses in many fisheries have encouraged substantial theoretical work on the challenging problem of optimal harvesting policy in response to demographic and environmental stochasticity (1–8). This has led to several sophisticated optimal harvest models supporting constant harvest mortality rates, threshold harvesting policies, or no-take reserves. Although these policies are sometimes feasible, in reality many management agencies have limited ability to control the number of resource users or harvest effort. This is particularly true of recreational harvesting, because of the open-access philosophy underlying sport fisheries and wildlife hunting. Even when harvest levels are directly set by regional managers, such control is often in the form of ad hoc quotas that vary from year to year. Modern harvesting theory is based on coupling harvest with dynamic variation in resource abundance. Here we show that weak compensatory response by harvesters or resource managers can itself generate

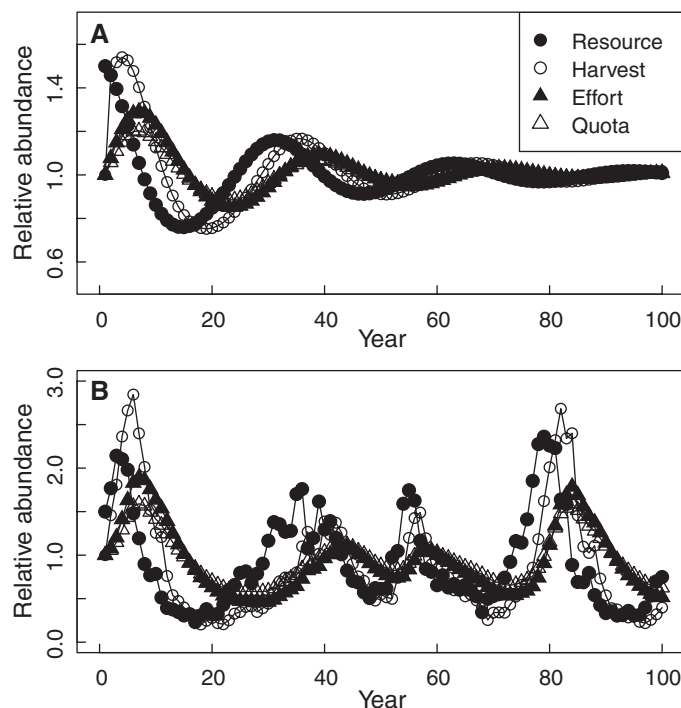
cyclic variation in resources, exacerbating the risk of collapse. Weak harvest regulation contributes to the problem rather than providing an acceptable management solution to resource fluctuation.

To consider this issue, we developed a dynamical system model in discrete time (9), based on simple intuitive assumptions about human

behavior, combined with mass action principles commonly applied in ecological models (10–16). The model assumes that harvesting is open access, meaning that there is no governmental restriction of harvest effort, but annual quotas are used to constrain the maximum harvest. Harvesters share information about their experiences during hunts or fishing trips, just as people talk about other important aspects of their lives. Mass action principles accordingly guide the sharing of social information, based on the probability of communication among networking members of a finite population of potential harvesters. Hence, we predict that the rate of change in effort from year to year should be a positive function of resource abundance but a negative function of current effort (9).

We assume that a similar mix of positive and negative feedbacks influences quota levels set by resource managers. High levels of resource abundance should encourage increasing quotas if managers are sensitive to meeting the rising expectations of harvesters, whereas declining resource levels should encourage the opposite response. Our model assumes that managers make such responses in incremental fashion, applying small percentage increases or decreases to the previous year's level in making their an-

Fig. 1. Predicted time series for the proposed dynamic harvest-effort-quota system, for a locally stable set of parameters ($\alpha = 0.3$, $K = 4$, $q = 0.0001$, $c = -0.1$, $w = 0.2$, $u = 0.00002$, $f = -0.05$, $i = 0.15$, $j = 0.1$) (9) (A) without any environmental stochasticity and (B) with the standard deviation in environmental stochasticity ($\epsilon = 0.20$). To simplify plotting on a single set of axes, variables were normalized by dividing yearly values by equilibrium values.



¹Department of Integrative Biology, University of Guelph, 50 Stone Road East, Guelph, Ontario N1G 2W1, Canada.

²Department of Ecology, Evolution, and Behavior, University of Minnesota, 1987 Upper Buford Circle, St. Paul, MN 55108, USA.

³Norwegian Institute for Nature Research, Tungasletta 2, N-7485 Trondheim, Norway. ⁴Centre for Conservation Biology, Department of Biology, Realfagbygget, NTNU, 7491 Trondheim, Norway.

*To whom correspondence should be addressed. E-mail: jfryxell@uoguelph.ca

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

The following resources related to this article are available online at www.sciencemag.org (this information is current as of May 13, 2010):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/328/5980/903>

Supporting Online Material can be found at:

<http://www.sciencemag.org/cgi/content/full/328/5980/903/DC1>

This article **cites 23 articles**, 4 of which can be accessed for free:

<http://www.sciencemag.org/cgi/content/full/328/5980/903#otherarticles>

This article appears in the following **subject collections**:

Ecology

<http://www.sciencemag.org/cgi/collection/ecology>

nual modification of quotas. Our model also assumes that managers are reluctant to exceed quota levels experienced in the past, leading to a negative feedback on rates of quota change (9).

These compensatory responses can be readily incorporated into a modified version of the Rosenzweig-MacArthur model of consumer-resource interactions (11, 12). We consider a discrete time formulation to be more realistic than a continuous formulation for several reasons: (i) most fish and wildlife species of interest have a single birth-pulse breeding season each year, (ii) harvesting is typically confined to a short season each year, and (iii) license sales and quota levels are usually set in advance of the harvest season. We accordingly structured our discrete time model by assuming that harvesting is followed by density-dependent recruitment according to the Ricker logistic formulation, with negative feedbacks on the exponential rate of population growth dictated by post-harvest resource abundance (9). For subsequent analyses of the deterministic form of the discrete time model, we re-scaled state variables to make them dimensionless. Local stability analysis (9) was used to compare the range of stable versus unstable parameter combinations, to determine whether damped oscillations characterize the response to local perturbations from equilibria and to estimate the period length of such damped oscillations (13–16).

In the absence of environmental stochasticity, our model predicts that equilibria are locally stable for a restricted set of parameter combinations, whereas stable limit cycles are the norm for other parameter combinations, a pattern often seen in ecological models (12–16). Deterministic instability is particularly likely when resource carrying capacity is high, the maximum rates of increase in effort or quotas are particularly high, or the rate of decline in effort or quotas at low resource abundance is particularly low (fig. S1). Even in stable systems, however, our model predicts that the approach to stability after perturbation will be extremely slow, characterized by damped oscillations over time (Fig. 1A). The return time for such damped cycles depends on intrinsic growth rates, but for many large-bodied species, including most mammals, birds, and sport fish, it may take decades for perturbed populations to recover their former abundance (fig. S2).

In the real world, populations are invariably subject to substantial levels of environmental and/or demographic stochasticity (17, 18). Damped oscillations in the deterministic system coupled with stochastic variation in population growth rates leads to quasi-cyclic fluctuation (15, 16) (Fig. 1B), with similar periodicity as that seen in simpler perturbed systems (Fig. 1A). In other words, dynamic variation in harvest effort and/or quotas, coupled with environmental or demographic stochasticity, should often lead to wildlife population cycles regardless of other ecological interactions.

Instability in our model traces from asynchrony in the dynamic responses of either harvesters or their regulating agencies to changes in resource

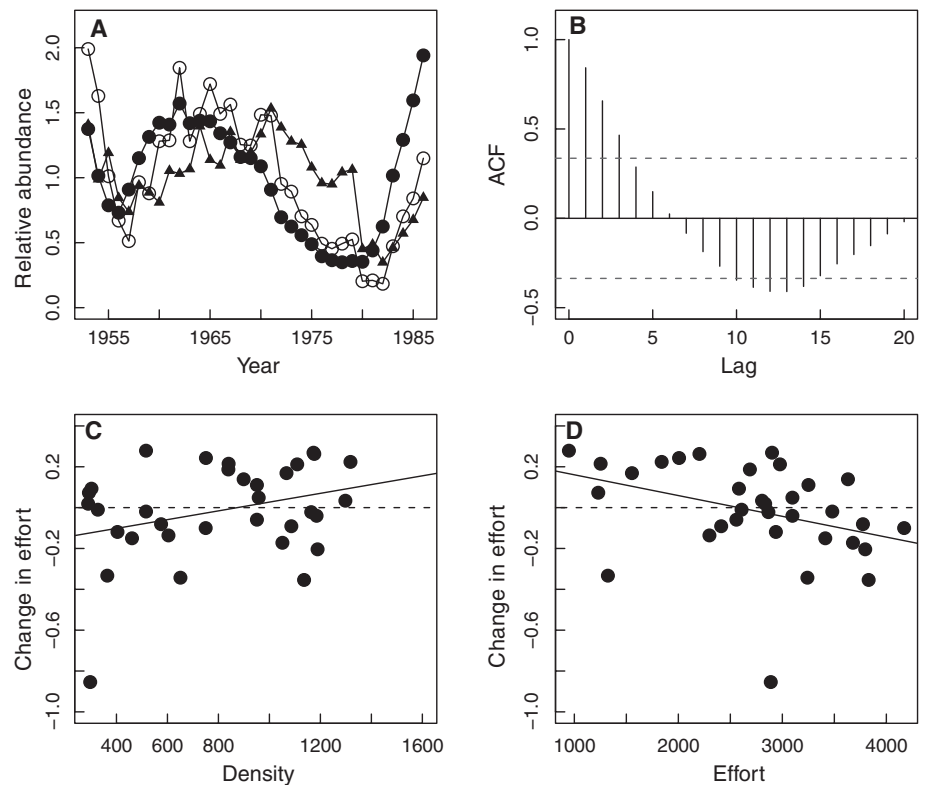


Fig. 2. Harvest data are illustrated for white-tailed deer from the Canonto District of Ontario, Canada. (A) Annual estimates of resource abundance (solid circles), harvest (open circles), and effort (solid triangles). To simplify plotting on a single set of axes, variables were normalized by dividing yearly values by the mean for the entire time series. Autocorrelation functions for resource abundance are shown in (B). All sample autocorrelation function values that exceed the horizontal bars are statistically significant ($P < 0.05$). The bottom two plots show rates of change in hunting effort (z) in relation to resource density (x) (C) and current effort (y) (D) ($z = 0.0706 + 0.000305x - 0.000104y$, $F_{2,30} = 5.51$, $P = 0.009$, $R^2 = 0.271$).

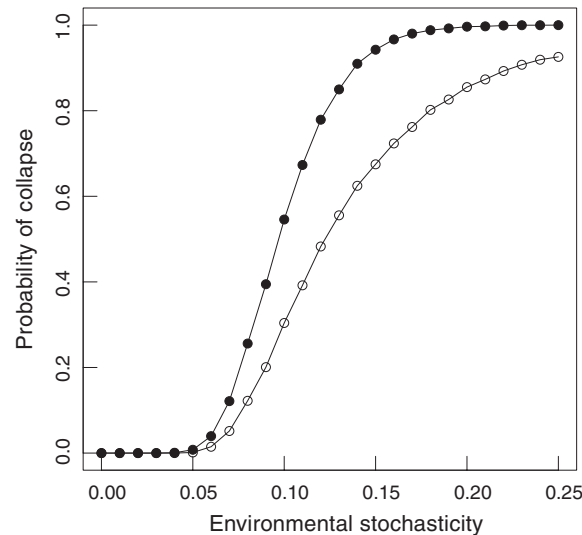
levels. Social adjustments in effort or regulatory adjustments in quotas occur at discrete, annual intervals because of short harvest season lengths. As a consequence, even locally stable systems would require a long time to equilibrate. In a realistically noisy world, such slow equilibration would translate into quasi-cycles of extended length, typically on the order of decades. The propensity for quasi-cyclic fluctuations is more pronounced for species with low intrinsic growth rates and modest levels of environmental stochasticity, such as large mammals, than it is for species with faster life histories or those subject to high variation in environmental conditions.

Our model suggests that harvested species are regulated by either harvester behavior or management policy, but not both. The structure of the model is such that either effort or quotas fully determine system equilibria, depending on which regulatory process supports the higher level of resources at equilibrium. Because both processes respond to changes in resource abundance, however, time dynamics of effort and quotas in unstable systems should become synchronized (Fig. 1B). Our model makes a number of testable predictions: (i) rates of change in effort and quotas should be positively associated with resource levels,

but negatively related to current levels of effort or quotas; (ii) harvested populations should often fluctuate with return times on the order of decades; and (iii) both effort and quotas should fluctuate in unison, but each should be far out of phase with resources, whereas harvest should have a lag of intermediate length.

As a proof of concept, we evaluated our model using three different populations of hunted ungulates (Canadian white-tailed deer and Norwegian moose) that had at least 20 years of continuous time series data available on population density, harvest, effort, and quota. Moose data for the Troms and Vest Agder Districts in Norway were provided by the National Moose-Monitoring Program (19). For the purposes of this paper, population abundance was estimated using moose seen per hunter day. Data on harvest level and hunter effort came from hunter questionnaires with high response rates. Data for white-tailed deer came from the Canonto District of Ontario. Deer harvests during 1953–1987 were limited to a 1- to 2-week period during the autumn. Access to the site was limited to two roads, both of which were closely monitored by wildlife personnel, who interviewed every hunter leaving the area to estimate hunt-

Fig. 3. Predicted probability of population collapse (to 10% of the equilibrium density) in a model system with constant effort (open circles) compared to one with weak compensatory changes in effort (solid circles), based on 5000 Monte Carlo replicates for each level of environmental stochasticity. Constant effort and quota levels were set at equilibria for the dynamic system. Simulations were based on the same locally stable set of parameters shown in Fig. 1 ($a = 0.3$, $K = 4$, $q = 0.0001$, $c = -0.1$, $w = 0.2$, $u = 0.00002$, $f = -0.05$, $i = 0.15$, $j = 0.1$) with the standard deviation in environmental stochasticity ϵ ranging between 0.00 and 0.25.



ing effort (days) and harvest. Deer population density in the 230-km² study area was estimated with virtual population analysis (20). For all study populations, exponential rates of increase in effort and quotas were estimated by $y_t = \ln(x_{t+1}/x_t)$, where y is the exponential rate of change, x is the response variable (either effort or quota), and t is time, measured in years. Linear regression was used to test whether rates of increase in effort and quotas were significantly related to resource abundance and current levels of either effort or quota. Standard time series analyses (autocorrelation and cross-correlation functions in the statistical software R, version 2.4) were applied to each data set to evaluate statistical support for recurrent versus damped oscillations and lag length for harvest, effort, and quotas relative to resource abundance (15). Although complete understanding of the population dynamics of long-lived species is enhanced by age-specific data on demographic rates (21), simpler nonstructured models, such as the Ricker equation used here, can often be used to capture the most salient dynamical features (15, 22). Simpler models have the added benefit of allowing the full range of analytical tools needed to evaluate the stability of complex dynamic systems.

All of the populations exhibited diagnostic patterns that were remarkably consistent with our model (Fig. 2 and figs. S3 and S4). Population abundance varied considerably over time, with the smallest range of variation recorded in moose in the Troms District (twofold) and the highest variability recorded in white-tailed deer (fivefold). In each population, autocorrelation functions provided evidence of phase-forgetting (damped) oscillations in resource abundance, with period lengths of 8, 24, and 30 years. Cross-correlation functions indicated a pronounced lag (2 to 6 years) between fluctuations in resource abundance and effort, a modest lag between harvest and effort, and little lag between quotas and effort (fig. S5). Changes in effort were pos-

itively related to resource abundance but negatively related to current levels of effort. Rates of change in quotas were also positively related to resource abundance but negatively related to quota level (fig. S6). All of these time series characteristics were predicted by the model.

Alternatively, changes in harvesting pressure can induce generational cycles in long-lived species by perturbations of the stable age distribution (21–23). We tested this possibility by time series analysis of the proportion of deer and moose young through time. Although strong cohorts were identified in all three study populations, autocorrelation functions showed no evidence of cyclic fluctuation in age structure that could explain the observed oscillations in overall population abundance (figs. S7 and S8). Thus, although it is theoretically possible that temporal variation in age structure can interact with harvest dynamics and population abundance to amplify unstable dynamics, generation cycles do not explain the observed oscillations in the abundance of white-tailed deer and moose.

Wildlife harvesting of species such as moose, elk, deer, bears, bobcats, and cougars is often conducted in a manner consistent with our model. In many jurisdictions, virtually anyone is eligible to apply for a hunting permit. There are often regional quotas on harvest, however, facilitated via telephone hotlines. Although there can be special restrictions on the age, sex, or location of harvested animals, regulatory procedures for a given species often tend to be fairly consistent across different wildlife agencies.

As a result of the open-access philosophy typically underlying wildlife harvest management (that is, employing few restrictions on participation in hunting, instead changing quotas on an annual basis), large-scale fluctuations in abundance are to be expected. A scattering of published papers in the literature are consistent with the unstable scenario we have drawn (1, 24–27), suggesting that quasi-periodic fluctuations in resources, exploitation rate, and quotas may be

much more common than previously realized. If we are correct in asserting that asynchronous dynamics of harvested resources, effort, and quotas often occur, due to delayed human behavioral responses to changing conditions, then conservation and management agencies might well benefit from more frequent reassessment of quota levels in response to undesirable changes in population levels occurring over the course of the hunting season, or from the institution of more conservative, temporally stable policies. Future harvest policies should aim to minimize variation in resource mortality risk by maintaining constant effort, setting quotas proportionate to current resource abundance, or shifting to threshold or proportional threshold approaches. Reconciling these alternative management options with open-access policies will no doubt prove challenging.

Weak harvest regulation should be particularly worrisome in populations with high levels of demographic or environmental stochasticity or with pronounced Allee effects due to predation, disease transmission, or reduced probability of breeding (28–30). Our simulations suggest that the risk of population collapse could be dramatically higher in systems with dynamic effort and quota levels (Fig. 3), simply because of extreme population excursions caused by quasi-periodic dynamics resulting from even mild levels of environmental stochasticity. When resource abundance is regulated by dynamic management responses via quotas, unstable systems are often vulnerable to extinction, even in the absence of Allee effects or stochastic variation in growth rates. These findings suggest that it is unwise to neglect dynamic patterns of change in both harvest effort and quotas in assessing long-term strategies for sustainable resource use.

References and Notes

1. B. Worm *et al.*, *Science* **325**, 578 (2009).
2. C. N. K. Anderson *et al.*, *Nature* **452**, 835 (2008).
3. J. R. Beddington, R. M. May, *Science* **197**, 463 (1977).
4. C. W. Clark, *Mathematical Bioeconomics: The Optimal Management of Renewable Resources* (Wiley, New York, 1990).
5. R. Lande, B.-E. Saether, S. Engen, *Ecology* **78**, 1341 (1997).
6. M. Mangel, *Ecol. Lett.* **1**, 87 (1998).
7. A. Hastings, L. W. Botsford, *Science* **284**, 1537 (1999).
8. M. G. Neubert, *Ecol. Lett.* **6**, 843 (2003).
9. Detailed materials and methods are provided in the supporting material on Science Online.
10. M. Schaefer, *Inter-Am. Trop. Tuna Comm. Bull.* **1**, 27 (1954).
11. M. L. Rosenzweig, R. H. MacArthur, *Am. Nat.* **97**, 209 (1963).
12. M. L. Rosenzweig, *Science* **171**, 385 (1971).
13. A. Hastings, *Population Biology: Concepts and Models* (Springer, New York, 1977).
14. S. P. Otto, T. Day, *A Biologist's Guide to Mathematical Modelling* (Princeton Univ. Press, Princeton, NJ, 2007).
15. P. Turchin, *Complex Population Dynamics* (Princeton Univ. Press, Princeton, NJ, 2003).
16. R. M. Nisbet, W. S. C. Gurney, *Modelling Fluctuating Populations* (Wiley, Chichester, UK, 1982).
17. M. Boyce, C. Haridas, C. Lee, *Trends Ecol. Evol.* **21**, 141 (2006).
18. R. Lande, S. Engen, B.-E. Saether, *Stochastic Population Dynamics in Ecology and Conservation* (Oxford Univ. Press, Oxford, 2003).
19. E. J. Solberg, B.-E. Saether, O. Strand, A. Loison, *J. Anim. Ecol.* **68**, 186 (1999).

20. J. M. Fryxell, D. J. T. Hussell, A. B. Lambert, P. C. Smith, *J. Wildl. Manage.* **55**, 377 (1991).
21. A. Beckerman, T. G. Benton, E. Ranta, V. Kaitala, P. Lundberg, *Trends Ecol. Evol.* **17**, 263 (2002).
22. T. Coulson *et al.*, *Ecology* **89**, 1661 (2008).
23. T. Coulson, F. Guinness, J. Pemberton, T. Clutton-Brock, *Ecology* **85**, 411 (2004).
24. G. Bell, P. Handford, C. Dietz, *J. Fish. Res. Board Can.* **34**, 842 (1977).
25. A. A. Berryman, *Oikos* **62**, 106 (1991).
26. R. McGarvey, *Can. J. Fish. Aquat. Sci.* **51**, 900 (1994).
27. C. Packer *et al.*, *PLoS ONE* **4**, e5941 (2009).
28. F. Courchamp, T. Clutton-Brock, B. Grenfell, *Trends Ecol. Evol.* **14**, 405 (1999).
29. P. A. Stephens, W. J. Sutherland, *Trends Ecol. Evol.* **14**, 401 (1999).
30. B. A. Melbourne, A. Hastings, *Nature* **454**, 100 (2008).
31. This work was supported by Discovery Grants to J.M.F. and K.M. from the Natural Sciences and Engineering Research Council of Canada, and NSF Grants for Long-Term Research in Environmental Biology and Biocomplexity to C.P. The contributions by B.E.S. and E.J.S. were funded by the Directorate for Nature Management and the Norwegian Research Council (Changing Landscapes, Lanskap i Endring, and project

196304/V40). We thank A. Fryxell, T. Avgar, M. Anderson, T. Nudds, and two anonymous reviewers for helpful suggestions. Details of the dynamical stability analysis are available on request from the senior author.

Supporting Online Material

www.sciencemag.org/cgi/content/full/328/5980/903/DC1
Materials and Methods

Figs. S1 to S8

Tables S1 to S5

References

10 December 2009; accepted 24 March 2010

10.1126/science.1185802

Molecular Identity of Dendritic Voltage-Gated Sodium Channels

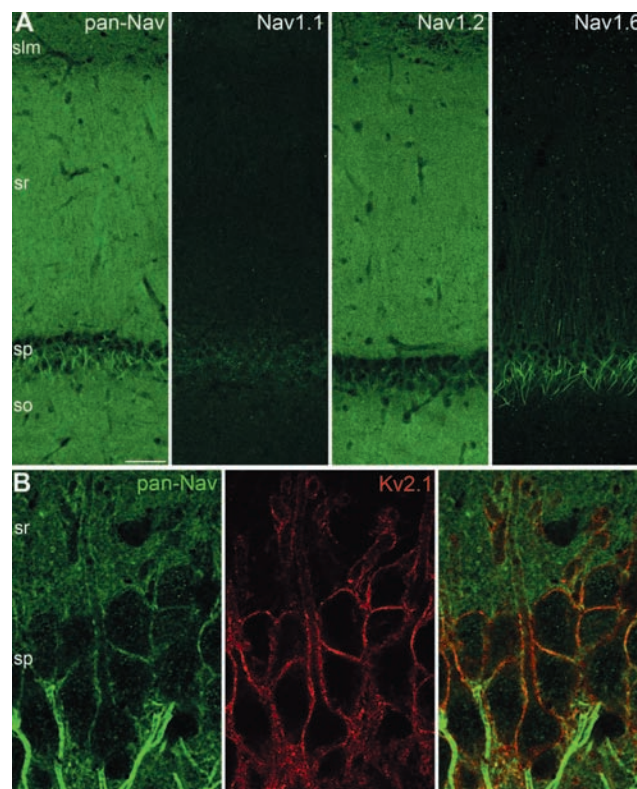
Andrea Lorincz* and Zoltan Nusser*

Active invasion of the dendritic tree by action potentials (APs) generated in the axon is essential for associative synaptic plasticity and neuronal ensemble formation. In cortical pyramidal cells (PCs), this AP back-propagation is supported by dendritic voltage-gated Na^+ (Nav) channels, whose molecular identity is unknown. Using a highly sensitive electron microscopic immunogold technique, we revealed the presence of the Nav1.6 subunit in hippocampal CA1 PC proximal and distal dendrites. Here, the subunit density is lower by a factor of 35 to 80 than that found in axon initial segments. A gradual decrease in Nav1.6 density along the proximodistal axis of the dendritic tree was also detected without any labeling in dendritic spines. Our results reveal the characteristic subcellular distribution of the Nav1.6 subunit, identifying this molecule as a key substrate enabling dendritic excitability.

Associative synaptic plasticity in cortical pyramidal cells (PC) is the most widely accepted cellular model of learning. An essential prerequisite of the model is that input synapses, which are distributed over an enormously large dendritic tree, must be capable of sensing the precise timing of the output signal. The most likely mechanism for this is the active invasion of the dendritic tree by fast sodium action potentials (APs) initiated in the axon initial segment (AIS). Modification of back-propagating APs therefore holds tremendous potential for altering synaptic plasticity and forming of neuronal representations. Voltage-gated Na^+ (Nav) currents have been detected in hippocampal and neocortical PC dendrites, where they not only support AP back-propagation but also underlie nonlinear synaptic integration and dendritic sodium spike generation [(1–6) and reviewed in (7–10)]. Patch-clamp experiments demonstrate that axonal and somatodendritic Nav currents differ in their activation and inactivation properties (11, 12), indicating either different Nav subunit compositions or distinct posttranslational modifications of identical subunits (12–14). However, the lack of subunit-specific drugs precludes the unequivocal identification of the subunit composition of the axosomatodendritic Nav channels using

functional approaches. Immunohistochemistry with subunit-specific antibodies offers an alternative experimental approach to address this issue. Nav1.1, Nav1.2, and Nav1.6 are the Nav subunits expressed in adult brains (15, 16). They have been

Fig. 1. Somatodendritic localization of voltage-gated sodium channels. (A) Immunofluorescence localization of the pan-Nav, Nav1.1, Nav1.2, and Nav1.6 subunits in the CA1 area. (B) Double immunofluorescence reaction shows weak pan-Nav immunolabeling along the Kv2.1 subunit immunoreactive somatodendritic plasma membrane of CA1 PCs. Note the much higher labeling intensity of the AISs. slm, stratum lacunosum-moleculare; sr, stratum radiatum; sp, stratum pyramidale; so, stratum oriens. Scale bars: (A), 50 μm ; (B), 10 μm .



Laboratory of Cellular Neurophysiology, Institute of Experimental Medicine, Hungarian Academy of Sciences, 1083 Budapest, Hungary.

*To whom correspondence should be addressed. E-mail: nusser@koki.hu (Z.N.); lorincz@koki.hu (A.L.)

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

The following resources related to this article are available online at www.sciencemag.org (this information is current as of May 13, 2010):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/328/5980/906>

Supporting Online Material can be found at:

<http://www.sciencemag.org/cgi/content/full/328/5980/906/DC1>

This article **cites 32 articles**, 15 of which can be accessed for free:

<http://www.sciencemag.org/cgi/content/full/328/5980/906#otherarticles>

This article appears in the following **subject collections**:

Neuroscience

<http://www.sciencemag.org/cgi/collection/neuroscience>

based fixation approach (see supporting online material) that vastly increased the strength of the reactions for almost all γ -aminobutyric acid type A (GABA_A) and glutamate receptors, and voltage-gated K⁺ and Na⁺ channels, and allowed the visualization of these proteins in postsynaptic membranes and AISs without antigen retrieval. Using this approach, we first employed pan-Nav antibodies to visualize all Nav subunits in the hippocampal CA1 area of the adult rat. Immunoreactions with pan-Nav antibodies revealed a very strong labeling of AISs and nodes of Ranvier in stratum pyramidale (SP) overlying a rather homogeneous staining of the strata oriens (SO), radiatum (SR) and lacunosum-moleculare (SLM) (Fig. 1A). The increased sensitivity afforded by the low-pH fixation condition allowed the visualization of pan-Nav immunosignal associated with somatic and proximal dendritic plasma membranes of PCs (Fig. 1B) that were also outlined by strong Kv2.1 subunit immunoreactivity. The intensity of the dendritic pan-Nav labeling was much weaker than that of the AISs, consistent with functional predictions (29). Interestingly, in some PCs, strongly pan-Nav-labeled AISs originated from the apical dendrites and a stronger pan-Nav labeling was associated with the plasma membrane of these

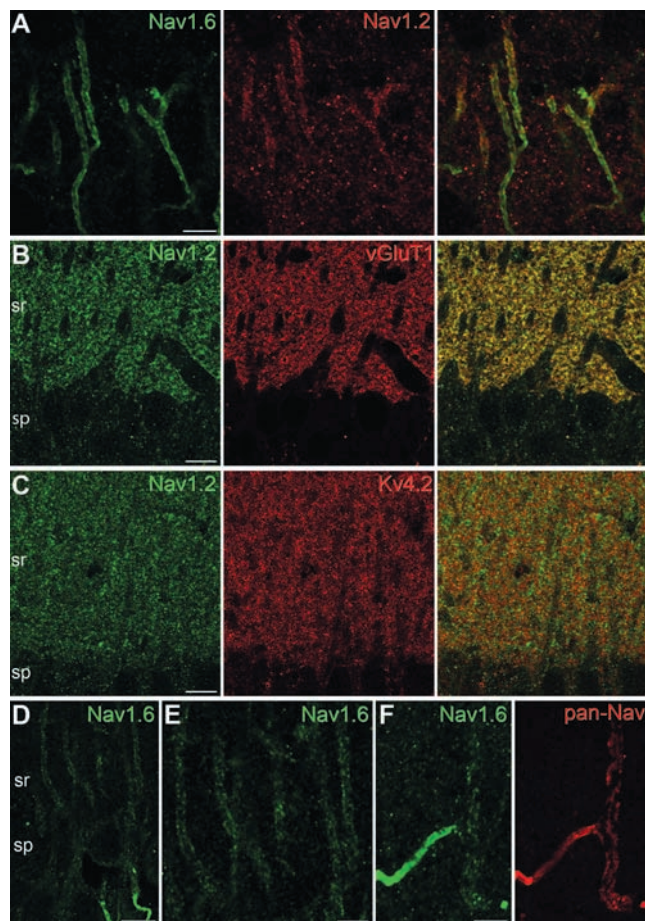
dendrites around the origin of the AIS (Fig. 2F and fig. S1). Our pan-Nav immunofluorescence reactions visualized Nav channels in the somatodendritic compartments of PCs and revealed large differences in Nav density between AISs and somatodendritic plasma membranes. However, the subunit composition, the presence or absence of Nav channels in small dendritic compartments such as small oblique dendrites and dendritic spines, and the exact quantitative difference in Nav density between distinct subcellular compartments remained elusive.

To address the first of these issues, we carried out immunofluorescence reactions with specific antibodies against the Nav1.1, Nav1.2, and Nav1.6 subunits. After confirming the specificity of the immunoreactions (figs. S2 and S3A), we investigated their distributions in the CA1 area. The labeling pattern obtained with the pan-Nav antibodies was essentially the sum of the patterns obtained for these three subunits (Fig. 1A). Immunosignal for the Nav1.1 subunit was restricted to small axonlike profiles in the SP; diffuse signal for the Nav1.2 subunit was observed in the neuropil of the SO, SR, and SLM; and intense Nav1.6 labeling was confined to AISs and nodes of Ranvier (Fig. 1A). Double immunolabeling for Kv3.1b and Nav1.1 subunits demonstrated that this Nav subunit is present

in small axonlike processes and AISs of Kv3.1b immunopositive GABAergic interneurons (fig. S3), but not in PCs, in agreement with a previous study (19). Double immunofluorescence reactions revealed the presence of the Nav1.2 subunit in the proximal part of PC AISs (Fig. 2A) (17), but no Nav1.2 subunit immunoreactivity could be detected in the plasma membranes of somata and proximal apical dendrites. The diffuse Nav1.2 subunit immunolabeling of the neuropil was associated with vGluT1 immunopositive presynaptic axons (Fig. 2B) but not with Kv4.2 immunoreactive dendritic shafts or spines of CA1 PCs (Fig. 2C), demonstrating the predominantly presynaptic expression of this subunit. Very intensely labeled AISs and nodes of Ranvier dominated the pictures obtained for the Nav1.6 subunit. In addition, in our best reactions, very weak plasma membrane-like labeling of PC apical dendrites could be observed (Fig. 2, D and E). This labeling decreased in intensity when individual dendrites were followed from their somata toward their apical tips (Fig. 1A). Weak dendritic Nav1.6 labeling could be detected in those PCs in which AISs originated from the apical dendrites (Fig. 2F). However, even in our best reactions, small-diameter oblique apical and basal dendrites, as well as dendritic spines, remained immunonegative. These results provide molecular identification of dendritic Nav channels and predict complex subcellular compartment- and distance-dependent regulation of the cell-surface density of Nav1.6 in PC dendritic trees.

To identify the potential presence of the Nav1.6 subunit in small dendritic compartments and to quantify the differences in the plasma membrane density of Nav1.6 in distinct subcellular compartments, we adopted SDS-FRL (23, 24), a recently developed quantitative EM immunolocalization method with exceptionally high sensitivity. The labeling efficiency for synaptic AMPA receptors of ~100% using SDS-FRL (30) is about 2.5 times as high as that achieved with a postembedding immunogold localization of the same subunits with an identical antibody (31). A labeling efficiency of 100% means that, on average, every protein is represented by a gold particle with a localization method that has a spatial resolution of ~25 nm. We therefore exploited the potential of this method for Nav1.6 localization by preparing carbon-platinum-carbon replicas of the CA1 area after high-pressure freezing and freeze-fracturing, and immunoreacted them with our antibodies to Nav1.6 (18). First we sought to examine the immunogold content of AISs (Fig. 3), the most intensely labeled subcellular structures found in the immunofluorescence reactions. Many elongated protoplasmic face (P-face) membrane structures in SP and SO contained a very high density of gold particles (Fig. 3A). These structures were then molecularly identified as AISs in double-labeling experiments with AIS markers such as pan-Neurofascin and Ankyrin-G (Fig. 3, B to D). The labeling on the P-face membranes is con-

Fig. 2. Immunofluorescence localization of the Nav1.2 and Nav1.6 subunits in CA1 PCs. (A) A double immunofluorescence reaction demonstrates the colocalization of the Nav1.6 and Nav1.2 subunits in the AISs of CA1 PCs. Immunolabeling for the Nav1.2 subunit is confined to the proximal part of the AISs. (B and C) Immunofluorescence double-labeling experiments reveal that the majority of the Nav1.2 subunit immunolabeling in stratum radiatum (sr) colocalizes with the presynaptic marker vGluT1 (B) but does not colocalize with the Kv4.2 subunit (C), a K⁺ channel subunit that is known to be present in PC dendritic shafts and spines. (D and E) An immunofluorescence reaction for the Nav1.6 subunit in the strata pyramidale (sp) and proximal radiatum (sr) of the CA1 area demonstrates weak plasma membrane-like labeling of PC proximal apical dendrites. Note the much higher labeling intensity of the AISs. (F) A double immunofluorescence reaction in the stratum radiatum shows the colocalization of pan-Nav (red) and the Nav1.6 subunit (green) in a strongly immunopositive AIS and its parent apical dendrite. Scale bars: (A), (E), and (F), 5 μ m; (B) to (D), 10 μ m.



sistent with the cytoplasmic location of the epitope recognized by our antibody (intracellular loop between domains II and III), given the accepted transmembrane topology of the Nav1.6 subunit. In a few cases, AISs were fractured together with a fragment of the corresponding somatic plasma membrane (Fig. 3H), allowing us to assess how Nav1.6 density changes in the AIS as a function of distance from the soma. The graded increase in the density of the Nav1.6 subunit along the proximodistal axis of PC AISs observed in the immunofluorescence reactions (Fig. 3G) could also be captured in replicas (Fig. 3, H to L). Furthermore, gold particles labeling the Nav1.6 subunit on AISs consistently avoided the perisynaptic and synaptic membrane areas of axo-axonic GABAergic synapses, identified by clusters of intramembrane particles (Fig. 3A and fig. S4B) or by intense clustering of gold particles labeling neuroligin-2 (fig. S4C) or GABA_A receptor $\beta 3$ subunit (fig. S4, D and E). The absence of the Nav1.6 subunit in GABAergic axo-axonic synapses is another feature that is reliably captured with SDS-FRL, as well as with double immunofluorescence labeling for Nav1.6 and, for example, the GABA_A receptor $\alpha 2$ subunit (fig. S4A). Large somatic P-face membrane structures in SP contained many more gold particles for the Nav1.6 subunit than membranes fractured to the extracellular face (E-face) (Fig. 4A). A similar picture was observed on proximal apical dendrites, where P-face dendritic shaft membranes contained gold particles (Fig. 4, B and C), consistent with our fluorescent observations. After the successful visualization of high-density Nav1.6 labeling in AISs and a much lower density in somata and proximal apical dendrites, we turned our attention to small dendritic compartments, which remained immunonegative even in our best fluorescence reactions. Dendritic shafts and spines of apical oblique dendrites were sampled at proximal (25 to 50 μm from SP) and distal (100 to 150 μm) locations in the SR, as well as dendritic shafts in SLM. Gold particles were observed on the P-face of dendritic shaft membranes in all locations, but dendritic spines were consistently immunonegative (Fig. 4, C and D). An intrinsic property of immunogold localization is that gold particles can be found in any location by chance. Therefore, the presence of Nav1.6 subunits cannot be concluded just from the presence of a few gold particles in a given compartment but requires a quantitative assessment of whether the particle density in a given location is significantly higher than that expected from a random process. We determined the nonspecific background labeling within our reactions by counting gold particles on the E-face plasma membranes (where the intracellular epitope of Nav subunits is not present) and found that on average 1.2 ± 0.4 gold particles ($n = 5$ rats) are expected to be present in every μm^2 membrane by chance (fig. S5). We then determined the density of gold particles labeling the Nav1.6 subunit on P-face plasma membranes of somata [6.3 ± 1.8 gold/ μm^2 (mean $\pm$ SD); $n =$

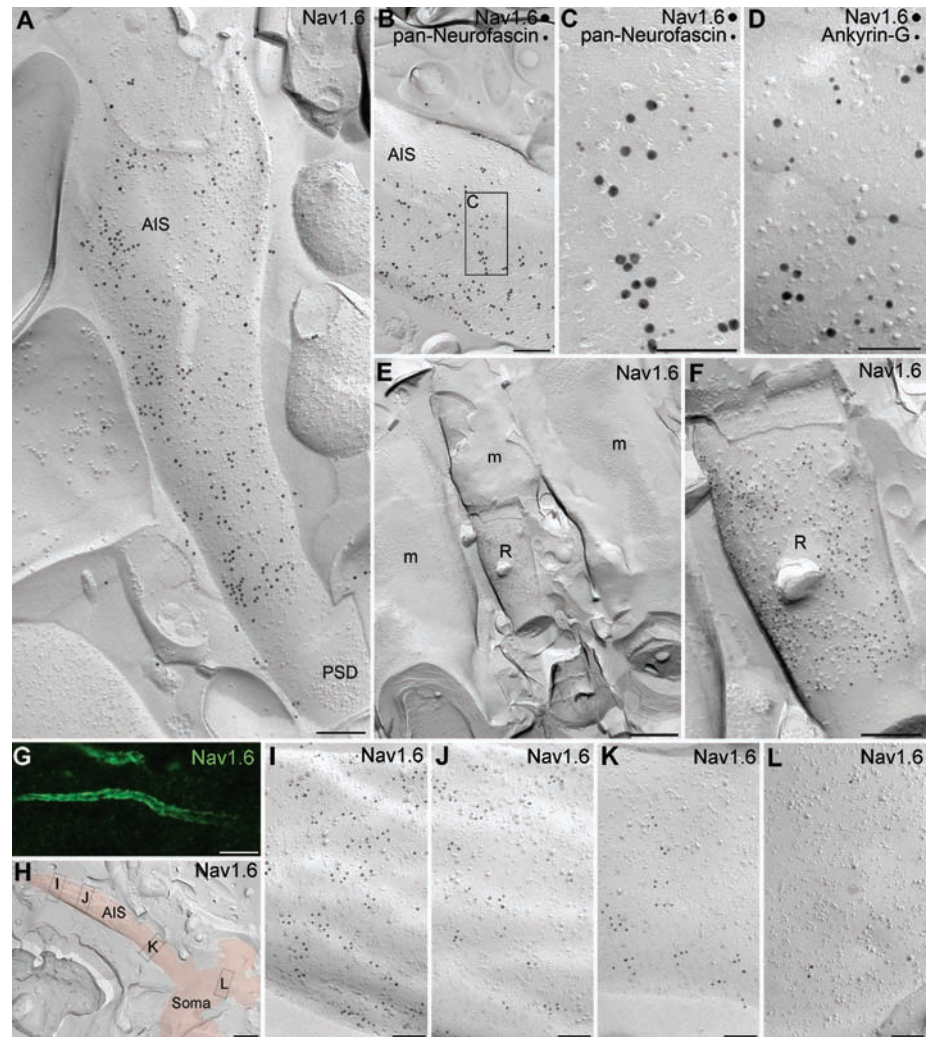
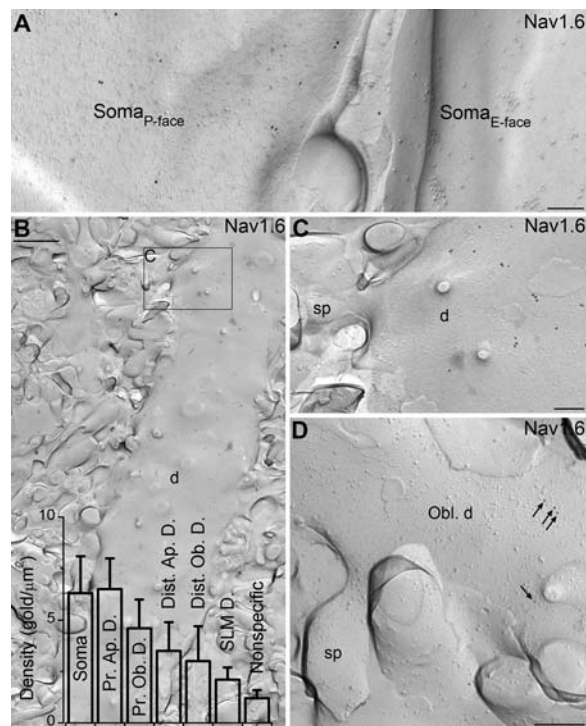


Fig. 3. High-resolution immunogold localization of the Nav1.6 subunit in AISs and nodes of Ranvier of CA1 PCs. (A) A high density of gold particles labeling the Nav1.6 subunit is found on the P-face of an AIS in the stratum pyramidale. The immunogold particles avoided the area of the postsynaptic density (PSD) of an axo-axonic synapse. (B) An AIS is labeled for the Nav1.6 subunit (15 nm gold) and the AIS marker pan-Neurofascin (10 nm gold). (C) A high-magnification view of the boxed area shown in (B). (D) High-magnification image of the P-face of an AIS labeled for the Nav1.6 subunit (15 nm gold) and Ankyrin-G (10 nm gold). (E) A low-magnification image shows a node of Ranvier (R) of a myelinated axon (m) in the alveus. (F) A high-magnification view of the node of Ranvier shown in (E). The P-face of the node of Ranvier membrane contains a high density of gold particles labeling the Nav1.6 subunit. Note the lack of labeling over the myelin [(E) and (F)]. (G) A gradual increase in the intensity of Nav1.6 immunofluorescence is found along the proximodistal axis of CA1 PC AISs. (H) A low-magnification image of a replica shows a fragment of a somatic plasma membrane and an emerging AIS (red). (I to L) High-magnification images of the boxed areas shown in (H). Images taken of the soma (L) and of the AIS at various distances from the soma [(I) to (K)] demonstrate an increase in the density of gold particles labeling the Nav1.6 subunit along the proximodistal axis of the AIS (I). Scale bars: (A), (B), and (F), 200 nm; (C), (D), and (I) to (L), 100 nm; (E), 500 nm; (G), 5 μm ; (H), 1 μm .

5 rats], proximal apical dendrites (6.5 ± 1.5 gold/ μm^2), proximal oblique dendrites (4.6 ± 1.4 gold/ μm^2), distal apical dendrites (3.5 ± 1.4 gold/ μm^2), distal oblique dendrites (3.0 ± 1.7 gold/ μm^2), and dendrites in SLM (2.1 ± 0.6 gold/ μm^2). Of these, somata, proximal apical, proximal oblique, and distal apical dendrites displayed a significantly higher density of gold particles than the background labeling (Fig. 4B) [one-way analysis of variance (ANOVA) with

Dunnett's post-hoc test, $P < 0.05$]. Thus, the exceptionally high sensitivity of the SDS-FRL method allowed us to reveal low, but significant, densities of the Nav1.6 subunit in these dendritic compartments, which remained immunonegative with the immunofluorescence method. Along these lines, it should be noted that the presence of low densities of other Nav subunits (e.g., Nav1.2) in somatodendritic compartments cannot be excluded based on immunofluorescence data only,

Fig. 4. Somatodendritic localization of the Nav1.6 subunit in CA1 PCs using SDS-FRL. **(A)** Low density of gold particles labeling the Nav1.6 subunit is found on the P-face of the somatic plasma membrane of a CA1 PC. No immunolabeling is seen on the E-face of a neighboring PC soma. **(B)** Low-magnification image of the P-face of a thick, spiny apical dendrite (d) of a CA1 PC in the proximal stratum radiatum labeled for the Nav1.6 subunit. Bar graphs show the densities of gold particles (mean $\pm$ SD) in somatodendritic subcellular compartments of CA1 PCs, some of which differ significantly from the nonspecific labeling (ANOVA: $P < 0.001$; Dunnett's post hoc test: soma, proximal apical dendrites, proximal oblique dendrites, distal apical dendrites; $P < 0.05$; $n = 5$ rats). **(C)** A high-magnification image of the boxed area shown in (B). Gold particles labeling the Nav1.6 subunit are distributed on the dendritic shaft (d) but avoid the dendritic spine (sp). **(D)** The P-face of a spiny oblique dendrite (Obl. d) in the proximal stratum radiatum contains a low density of gold particles for the Nav1.6 subunit (arrows), with no particle in the spine (sp). Scale bars: (B), 1 μ m; (A), (C), and (D), 200 nm.



but highly sensitive reactions must be accomplished with SDS-FRL to unequivocally determine their presence or absence in identified subcellular compartments. Finally, we compared the Nav1.6 subunit density in somatodendritic compartments to that found in AISs and nodes of Ranvier. Nodes of Ranvier in the axons were very strongly immunolabeled for the Nav1.6 subunit (Fig. 3, E and F), containing immunogold particles at almost twice the density of that observed in AISs (346.3 ± 82.9 versus 186.9 ± 29.9 gold/ μ m²). The AISs contained densities of gold particles that were 39 ± 9 and 36 ± 6 times as high as those in somata and proximal apical dendrites, respectively. These ratios are very similar to those estimated (29) for neocortical layer 5 PCs using whole-cell Nav current recordings (a factor of ~ 45) and Na^+ imaging (a factor of ~ 30). When dendritic compartments at the same distance from the soma were compared, we found that oblique dendrites contain less Nav1.6 subunit than main apical dendrites. There was also a clear tendency for a distance-dependent reduction in the density of gold particles along the proximodistal axis of the dendritic tree (Fig. 4B). We could not detect immunogold particles labeling the Nav1.6 subunit above the background level in preterminal axons, axon terminals, or dendritic spines.

Using an improved immunofluorescence method and SDS-FRL, we identify Nav1.6 as the prominent somatodendritic Nav channel subunit in hippocampal PCs. Using SDS-FRL,

we demonstrate a characteristic subcellular distribution pattern of the Nav1.6 subunit on the axosomatodendritic compartments of PCs, revealing a very dramatic drop (by a factor of 40 to 70) in density from nodes of Ranvier and AISs to somata, and a distance-dependent decrease in density along the proximodistal axis of the dendritic tree. By exploiting the exquisite sensitivity of the SDS-FRL method, we obtained a gold particle density for the Nav1.6 subunit of 187 ± 30 (gold/ μ m²) in AISs of CA1 PCs. A previous study (29) using in vitro electrophysiology, Na^+ imaging, and multicompartmental modeling estimated a conductance density of Nav channels in neocortical layer 5 PC AISs of $2500 \text{ pS}/\mu\text{m}^2$; given a 17-pS single-channel conductance, this predicts a functional channel density of ~ 150 channels per μm^2 . If CA1 PCs have an AIS Nav channel density similar to layer 5 PCs, we can conclude that the labeling efficiency of our SDS-FRL method is close to 100%, similar to that obtained for synaptic AMPA receptors (30). Given this estimated labeling efficiency, our results predict a Nav channel density (channels per μm^2 plasma membrane) of 5 in somata and proximal apical dendrites, 3 in proximal oblique dendrites, and 2 in distal apical dendrites. The differences in the density of Nav1.6 in somatic and dendritic compartments represent only one level of complexity. Additional compartment-dependent variations in, for example, the phosphorylation states of the channels (12) could further increase functional diversity. Our results demonstrate that Nav1.6 is

the main Nav subunit in the somatodendritic compartments of CA1 PCs, whereas both Nav1.6 and Nav1.2 subunits are present in AISs (17). It remains to be determined whether the axosomatic difference in the subunit composition of Nav channels is solely responsible for the reported differences in activation threshold or whether additional differential regulation of axonal versus somatic Nav channels also exists. Irrespective of that, our data support the view that the large axonal Nav channel density is primarily responsible for the fact that the axon has the lowest voltage threshold for generating action potentials (32).

References and Notes

1. A. Losonczy, J. K. Makara, J. C. Magee, *Nature* **452**, 436 (2008).
2. M. E. Larkum, J. Waters, B. Sakmann, F. Helmchen, *J. Neurosci.* **27**, 8999 (2007).
3. M. E. Larkum, T. Nevian, M. Sandler, A. Polsky, J. Schiller, *Science* **325**, 756 (2009).
4. A. Kamondi, L. Acsády, G. Buzsáki, *J. Neurosci.* **18**, 3919 (1998).
5. A. Losonczy, J. C. Magee, *Neuron* **50**, 291 (2006).
6. S. Gasparini, J. C. Magee, *J. Neurosci.* **26**, 2088 (2006).
7. M. Migliore, G. M. Shepherd, *Nat. Rev. Neurosci.* **3**, 362 (2002).
8. D. Johnston, J. C. Magee, C. M. Colbert, B. R. Christie, *Annu. Rev. Neurosci.* **19**, 165 (1996).
9. J. Magee, D. Hoffman, C. Colbert, D. Johnston, *Annu. Rev. Physiol.* **60**, 327 (1998).
10. G. Stuart, N. Spruston, B. Sakmann, M. Häusser, *Trends Neurosci.* **20**, 125 (1997).
11. C. M. Colbert, J. C. Magee, D. A. Hoffman, D. Johnston, *J. Neurosci.* **17**, 6512 (1997).
12. S. Gasparini, J. C. Magee, *J. Physiol.* **541**, 665 (2002).
13. J. Xu, N. Kang, L. Jiang, M. Nedergaard, J. Kang, *J. Neurosci.* **25**, 1750 (2005).
14. C. M. Colbert, D. Johnston, *J. Neurophysiol.* **79**, 491 (1998).
15. S. Beckh, M. Noda, H. Lübbert, S. Numa, *EMBO J.* **8**, 3611 (1989).
16. J. S. Trimmer, K. J. Rhodes, *Annu. Rev. Physiol.* **66**, 477 (2004).
17. W. Hu et al., *Nat. Neurosci.* **12**, 996 (2009).
18. A. Lorincz, Z. Nusser, *J. Neurosci.* **28**, 14329 (2008).
19. I. Ogiwara et al., *J. Neurosci.* **27**, 5903 (2007).
20. T. Boiko et al., *J. Neurosci.* **23**, 2306 (2003).
21. A. Van Wart, J. S. Trimmer, G. Matthews, *J. Comp. Neurol.* **500**, 339 (2007).
22. B. Gong, K. J. Rhodes, Z. Bekele-Arcuri, J. S. Trimmer, *J. Comp. Neurol.* **412**, 342 (1999).
23. K. Fujimoto, *J. Cell Sci.* **108**, 3443 (1995).
24. J. E. Rash, T. Yasumura, C. S. Hudson, P. Agre, S. Nielsen, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 11981 (1998).
25. M. Masugi-Tokita, R. Shigemoto, *Curr. Opin. Neurobiol.* **17**, 387 (2007).
26. A. Lorincz, Z. Nusser, *J. Neurosci.* **28**, 9083 (2008).
27. J.-M. Fritschy, O. Weinmann, A. Wenzel, D. Benke, *J. Comp. Neurol.* **390**, 194 (1998).
28. M. Watanabe et al., *Eur. J. Neurosci.* **10**, 478 (1998).
29. M. H. Kole et al., *Nat. Neurosci.* **11**, 178 (2008).
30. J. Tanaka et al., *J. Neurosci.* **25**, 799 (2005).
31. Z. Nusser et al., *Neuron* **21**, 545 (1998).
32. M. H. Kole, G. J. Stuart, *Nat. Neurosci.* **11**, 1253 (2008).
33. A.L. is the recipient of a János Bolyai Scholarship of the Hungarian Academy of Sciences. Z.N. is the recipient of a European Young Investigator Award (www.esf.org/euryi) and Wellcome Trust Project and Equipment grants. We thank D. Rónaszéki and É. Dobai for their excellent technical help, W. Sieghart and J.-M. Fritschy for antibodies to the GABA_A receptors, and M. Eyre and M. Farrant for their comments on the manuscript.

Supporting Online Material

www.sciencemag.org/cgi/content/full/328/5980/906/DC1
Materials and Methods
Figs. S1 to S5

4 February 2010; accepted 2 April 2010
10.1126/science.1187958

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

The following resources related to this article are available online at www.sciencemag.org (this information is current as of May 13, 2010):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:
<http://www.sciencemag.org/cgi/content/full/328/5980/910>

Supporting Online Material can be found at:
<http://www.sciencemag.org/cgi/content/full/328/5980/910/DC1>

This article **cites 22 articles**, 8 of which can be accessed for free:
<http://www.sciencemag.org/cgi/content/full/328/5980/910#otherarticles>

This article appears in the following **subject collections**:
Microbiology
<http://www.sciencemag.org/cgi/collection/microbio>

A Plant-Like Kinase in *Plasmodium falciparum* Regulates Parasite Egress from Erythrocytes

Jeffrey D. Dvorin,^{1,2} Derek C. Martyn,⁴ Saurabh D. Patel,^{1,3} Joshua S. Grimley,^{5*} Christine R. Collins,⁶ Christine S. Hopp,⁷ A. Taylor Bright,⁸ Scott Westenberger,^{8†} Elizabeth Winzeler,^{8,9} Michael J. Blackman,⁶ David A. Baker,⁷ Thomas J. Wandless,⁵ Manoj T. Duraisingh^{1‡}

Clinical malaria is associated with the proliferation of *Plasmodium* parasites in human erythrocytes. The coordinated processes of parasite egress from and invasion into erythrocytes are rapid and tightly regulated. We have found that the plant-like calcium-dependent protein kinase PfCDPK5, which is expressed in invasive merozoite forms of *Plasmodium falciparum*, was critical for egress. Parasites deficient in PfCDPK5 arrested as mature schizonts with intact membranes, despite normal maturation of egress proteases and invasion ligands. Merozoites physically released from stalled schizonts were capable of invading new erythrocytes, separating the pathways of egress and invasion. The arrest was downstream of cyclic guanosine monophosphate-dependent protein kinase (PfPKG) function and independent of protease processing. Thus, PfCDPK5 plays an essential role during the blood stage of malaria replication.

Plasmodium falciparum malaria remains a leading cause of morbidity and mortality worldwide, and resistance to existing antimalarials is a growing problem. Signal transduction plays a major role in key transitions of the parasite life cycle. Calcium is thought to be a regulator of parasite egress from and invasion into host cells. Multiple calcium-dependent protein kinases (CDPKs), a subfamily absent from the human genome, have been identified in apicomplexan organisms (1–8).

The *PfCDPK5* gene (PF13_0211) is transcribed in mature blood-stage schizonts and invasive merozoites (9), suggesting a role in egress and/or erythrocyte invasion (fig. S1). PfCDPK5 exhibits a canonical multidomain structure with a serine/threonine kinase domain followed by a C-terminal calmodulin-like domain comprising four Ca^{2+} -binding EF hands (Fig. 1A) (10). This structure is predicted to facilitate rapid kinase activation after locally increased Ca^{2+} concentration. Recombinant PfCDPK5 phosphorylated an artificial target protein in the presence of Ca^{2+} , and this was inhibited by the calcium chelator EGTA, demonstrating that PfCDPK5

is a bona fide CDPK (Fig. 1B) (11). In the absence of substrate, recombinant PfCDPK5 autophosphorylated.

We used the destabilizing domain (DD) system to regulate the level of PfCDPK5 expression in *P. falciparum* (12–15). DD-fusion proteins are expected to be rapidly degraded in the absence of the ligand Shield-1 (Shld1) and

stabilized in its presence (Fig. 1C). We genetically fused DD_{TM}, a strongly destabilizing DD derivative (16) to the C terminus of PfCDPK5, generating the D10-PfCDPK5-DD_{TM} and D10-PfCDPK5-HA-DD_{TM} parasite lines (fig. S2), which expressed PfCDPK5 in schizonts (Fig. 1D). PfCDPK5 has potential palmitoylation sites but no myristoylation site like its paralog PfCDPK1 (17). Carbonate extraction, which separates soluble from membrane-associated proteins, showed that PfCDPK5 was associated with parasite membranes (Fig. 1E). This may be important for maintaining the kinase in close proximity to its substrate(s).

Newly invaded ring-stage D10-PfCDPK5-DD_{TM} and D10-PfCDPK5-HA-DD_{TM} parasites were maintained in the presence or absence of Shld1 until the mature schizont stage. PfCDPK5 levels were decreased ~60 to 80% in the absence of Shld1 (Fig. 1F) in the detergent-insoluble fraction. To evaluate the requirement for PfCDPK5 for *P. falciparum* growth, we examined replication rates of D10-PfCDPK5-DD_{TM} parasites in the presence and absence of Shld1 (Fig. 1G). As a control, we created parasites with DD_{TM} fused to PfCDPK4 (D10-PfCDPK4-DD_{TM}), a paralog not essential for asexual replication (1, 6). Without Shld1, the D10-PfCDPK5-DD_{TM} parasites failed to proliferate, demonstrating that PfCDPK5 is essential. In contrast, D10-PfCDPK4-DD_{TM} parasites grew normally in the absence of Shld1, despite a >90% drop in PfCDPK4 levels (fig.

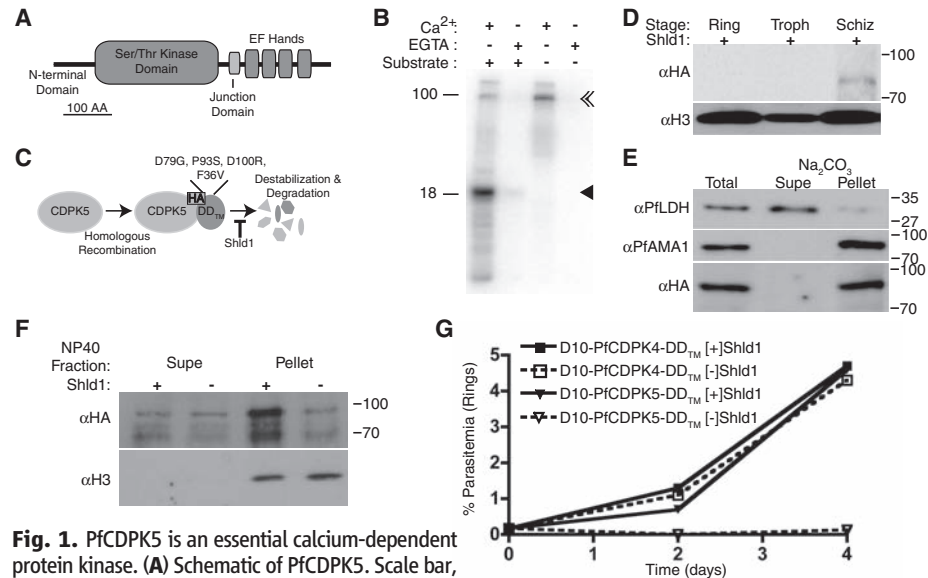


Fig. 1. PfCDPK5 is an essential calcium-dependent protein kinase. (A) Schematic of PfCDPK5. Scale bar, 100 amino acids. (B) Recombinant GST-PfCDPK5-6His (94 kD) was incubated with or without substrate (myelin basic protein, 18 kD), with 1.1 mM Ca^{2+} or 1 mM EGTA, and ^{32}P -g-ATP. In the presence of Ca^{2+} , PfCDPK5 phosphorylates itself (double arrowhead) and substrate (arrowhead). (C) PfCDPK5 fused to DD_{TM} is targeted for degradation but is stabilized by Shld1. (D) Protein lysates from D10-PfCDPK5-HA-DD_{TM} ring (0 to 20 hours), trophozoite (20 to 36 hours), and schizont-stage (36 to 48 hours) parasites cultured with Shld1 and probed with antibody to hemagglutinin (HA) (PfCDPK5), or antibody to histone H3 (loading control). (E) Na_2CO_3 -extracted schizonts were probed with antibody to PfLDH (cytoplasmic fraction), antibody to PfAMA1 (membrane fraction), or antibody to HA. (F) D10-PfCDPK5-HA-DD_{TM} parasites were grown [–] or [–] Shld1 until 44 hours, incubated with E-64 for 10 hours, harvested, fractionated with 0.6% NP-40, and probed with antibody to HA and antibody to histone H3. (G) Representative replication curves. D10-PfCDPK5-DD_{TM} and D10-PfCDPK4-DD_{TM} parasites were cultured [–] and [–] Shld1.

¹Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, MA 02115, USA. ²Division of Infectious Diseases, Children's Hospital Boston, Boston, MA 02115, USA. ³Division of Gastroenterology, Children's Hospital Boston, Boston, MA 02115, USA. ⁴The Broad Institute, Cambridge, MA 02142, USA. ⁵Department of Chemical and Systems Biology, Stanford University, Stanford, CA 94305, USA. ⁶Division of Parasitology, MRC National Institute for Medical Research, London NW7 1AA, UK. ⁷London School of Hygiene and Tropical Medicine, London, WC1E 7HT, UK. ⁸Department of Cell Biology, The Scripps Research Institute, La Jolla, CA 92037, USA. ⁹Genomics Institute of the Novartis Research Foundation, San Diego, CA 92121, USA.

*Present address: Department of Biochemistry, Duke University Medical Center, Durham, NC 27710, USA.

†Present address: Illumina, Inc., San Diego, CA 92121, USA.

‡To whom correspondence should be addressed. E-mail: mduraisi@hsph.harvard.edu

S3). D10-PfCDPK5-DD_{TM} parasitemia levels remained <0.2% for >10 days after Shld1 removal. However, after 10 to 12 days, a Shld1-independent revertant population emerged (D10-PfCDPK5-DD_{TM}-Rev) (fig. S4A), with a

concomitant genomic alteration at the *PfCDPK5* locus (fig. S2B). We generated PfCDPK5-deficient parasites in the 3D7 strain, demonstrating a similar arrest in proliferation in a different strain of *P. falciparum* (fig. S4B).

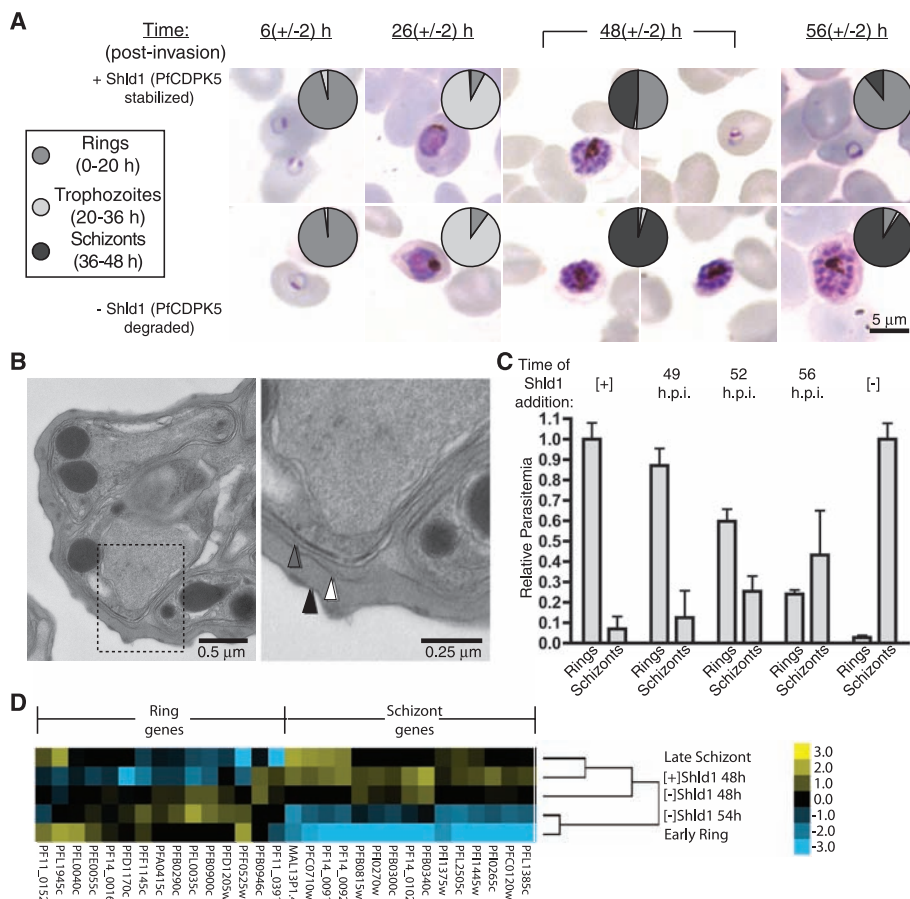


Fig. 2. PfCDPK5-deficient parasite arrest before egress. (A) Light microscopy of

synchronized D10-PfCDPK5-DD_{TM} parasites grown [+] or [-] Shld1 at indicated h.p.i. Pie charts show relative proportions of rings, trophozoites, and schizonts ($N = 100$). (B) D10-PfCDPK5-DD_{TM} parasites were grown [-]Shld1 until 56 h.p.i. and visualized by electron microscopy. The erythrocyte PM (black triangle), PVM (white triangle), and the parasite PM (gray triangle) were intact. (C) Ring-stage D10-PfCDPK5-DD_{TM} parasites were grown [+] or [-] Shld1 until 48 h.p.i. In parallel [-]Shld1 cultures, Shld1 was added back at the indicated times (49, 52, and 56 h.p.i.). Rings and schizonts were counted 4 hours later (mean $\pm$ SD; $N = 4$). (D) RNA expression patterns from 3D7-PfCDPK5-DD_{TM} parasites at 48 h.p.i. [+] and [-] Shld1 and 54 h.p.i. [-]Shld1. Log₂-transformed expression pattern shown for 15 late schizont-stage and 15 early ring-stage genes.

Morphological analysis over the 48-hour blood-stage asexual cycle (Fig. 2A) revealed healthy rings 6 ± 2 hours post-invasion (h.p.i.) and trophozoites 26 ± 2 h.p.i., irrespective of the presence of Shld1. At 48 ± 2 h.p.i., schizonts were observed both with [+] and without [-] Shld1. New rings were observed in the presence of Shld1 but were >95% reduced in the [-]Shld1 cultures. Eight hours later, the majority of [+]Shld1 parasites had reinvaded to form rings, but the [-]Shld1 parasites, deficient for PfCDPK5, remained stalled as late schizonts, with a >90% reduction in new ring formation. Thus, PfCDPK5 plays an essential role in parasite proliferation and egress from the erythrocyte. Functional knockout of PfCDPK5 did not affect the number of merozoites per schizont (fig. S5). Arrested schizonts were ultrastructurally normal, and the erythrocyte plasma membrane (PM), parasitophorous vacuole membrane (PVM), and parasite PM were intact (Fig. 2B). The block thus occurred after schizont differentiation but before PVM rupture.

To determine whether PfCDPK5-deficient schizonts had undergone aberrant differentiation or were irreversibly prevented from egress and/or invasion, we performed Shld1 rescue experiments (Fig. 2C). Upon observation of reinvaded rings in the [+]Shld1 parasites (48 h.p.i.), we allowed parallel [-]Shld1 parasites to incubate for an additional 1, 4, or 8 hours before adding back Shld1; new ring formation and levels of nonruptured schizonts were measured 4 hours after the addition of Shld1. Newly invaded rings were observed, albeit at lower levels, even in parasites that had been PfCDPK5-deficient for 52 h.p.i. and 56 h.p.i. Thus, PfCDPK5-deficient arrested schizonts had differentiated normally and retained the ability to egress and reinvade. Gene expression profiles in PfCDPK5-deficient parasites at 46 ± 2 h.p.i. correlated most closely with the expression pattern for late schizonts ($r = 0.83$) (9). Focusing on a subset of 30 differentially expressed genes between late schizonts and early rings, the expression pattern of the arrested parasites progressed toward the early ring phenotype (Fig. 2D and data file S1).

Recent studies have identified a parasite protease cascade required for egress from erythrocytes (18, 19) and maturation of merozoite surface proteins. PfSUB1, released from specialized secretory organelles called exonemes, cleaves several proteins, including the protease PfSERA5 and the invasion ligand merozoite surface protein 1 (PfMSP1) (19, 20). In our egress-blocked parasites, processing of PfSUB1 (mature form, ~47 kD), PfSERA5 (mature form, ~50 kD), and PfMSP1 (mature form, ~42 kD) were all unaffected (Fig. 3A), and in the PfCDPK5-deficient arrested parasites, PfSUB1 discharge and enzyme activity were unaffected. Although PfSERA5 processing is necessary for parasite egress, it is not sufficient. PfCDPK5 mediates a separate and downstream signal to the protease cascade required for parasite egress.

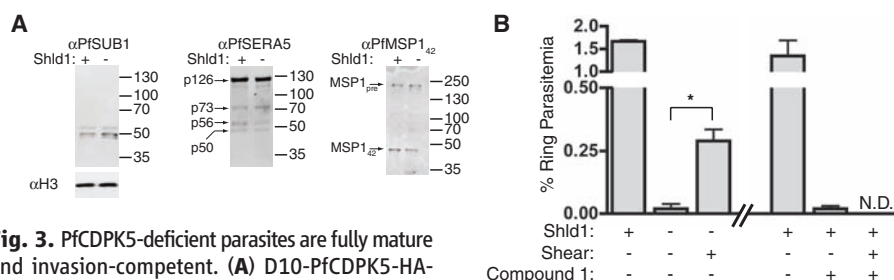


Fig. 3. PfCDPK5-deficient parasites are fully mature and invasion-competent. (A) D10-PfCDPK5-HA-DD_{TM} parasites were grown [+] and [-] Shld1, harvested by Percoll-purification at 48 h.p.i., and analyzed by Western blot with antibody to PfSUB1, antibody to PfSERA5, and antibody to PfMSP1₄₂. (B) D10-PfCDPK5-DD_{TM} parasites were grown until 48 h.p.i. ([+]/[-] 0.5 μM Shld1, [-]/[-] 4 μM compound 1). Subsets of parasites were sheared through a 28 G needle. Ring parasitemia was determined 4 hours later. Mean $\pm$ SD (or data range); $N = 2$ to 7; *, $P < 0.05$, unpaired t test.

Selective inhibition of the cyclic guanosine monophosphate (cGMP)-dependent kinase PfPKG with 4-[2-(fluorophenyl)-5-(1-methylpiperidine-4-yl)-1*H* pyrrol-3-yl]pyridine (compound 1) blocks egress (21). Treatment of schizonts with compound 1 inhibited processing of PfMSP1 by PfSUB1 (fig. S6A). However, compound 1 did not inhibit PfSUB1 enzyme activity directly (fig. S6B), suggesting that compound 1 treatment interferes with access of PfSUB1 to its endogenous substrates, perhaps by preventing exome discharge.

To determine whether PfCDPK5-deficient merozoites are capable of erythrocyte invasion, we mechanically disrupted arrested schizonts by needle-shearing. Newly invaded rings were observed at a significantly higher rate in the sheared cultures compared with unshaded controls (Fig. 3B). Thus, the egress trigger may be distinct from the erythrocyte invasion trigger. Isolation of viable PfCDPK5-deficient merozoites provides a valuable source for studies of erythrocyte invasion. In contrast to the invasion-competent PfCDPK5-deficient merozoites, mechanically disrupted compound 1-treated parasites did not generate newly invaded rings (Fig. 3B), consistent with these parasites being blocked at an earlier step.

We propose a modified model of *P. falciparum* egress from erythrocytes (fig. S7). The protease

cascade is necessary but not sufficient for parasite egress. After an egress trigger, a transient increase in Ca^{2+} concentration activates PfCDPK5 to amplify and transmit the signal for parasite egress. The final stages of PVM and erythrocyte PM rupture probably involve molecules downstream of PfCDPK5 activation, such as perforin-like proteins, membrane channels, and/or proteases (22, 23).

References and Notes

- O. Billker *et al.*, *Cell* **117**, 503 (2004).
- A. Coppi *et al.*, *Cell Host Microbe* **2**, 316 (2007).
- P. M. Färber, R. Graeser, R. M. Franklin, B. Kappes, *Mol. Biochem. Parasitol.* **87**, 211 (1997).
- J. L. Green *et al.*, *J. Biol. Chem.* **283**, 30980 (2008).
- T. Ishino, Y. Orito, Y. Chinzei, M. Yuda, *Mol. Microbiol.* **59**, 1175 (2006).
- N. Kato *et al.*, *Nat. Chem. Biol.* **4**, 347 (2008).
- H. Kieschnick, T. Wakefield, C. A. Narducci, C. Beckers, *J. Biol. Chem.* **276**, 12369 (2001).
- Y. Zhao, B. Kappes, R. M. Franklin, *J. Biol. Chem.* **268**, 4347 (1993).
- K. G. Le Roch *et al.*, *Science* **301**, 1503 (2003).
- J. F. Harper, A. Harmon, *Nat. Rev. Mol. Cell Biol.* **6**, 555 (2005).
- Materials and methods are available as supporting material on Science Online.
- C. M. Armstrong, D. E. Goldberg, *Nat. Methods* **4**, 1007 (2007).
- L. A. Banaszynski, L.-C. Chen, L. A. Maynard-Smith, A. G. L. Ooi, T. J. Wandless, *Cell* **126**, 995 (2006).
- A. Herm-Götz *et al.*, *Nat. Methods* **4**, 1003 (2007).
- I. Russo, A. Oksman, B. Vaupel, D. E. Goldberg, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 1554 (2009).
- B. W. Chu, L. A. Banaszynski, L.-C. Chen, T. J. Wandless, *Bioorg. Med. Chem. Lett.* **18**, 5941 (2008).
- C. Möskes *et al.*, *Mol. Microbiol.* **54**, 676 (2004).
- S. Arastu-Kapur *et al.*, *Nat. Chem. Biol.* **4**, 203 (2008).
- S. Yeoh *et al.*, *Cell* **131**, 1072 (2007).
- K. Koussis *et al.*, *EMBO J.* **28**, 725 (2009).
- H. M. Taylor *et al.*, *Eukaryot. Cell* **9**, 37 (2010).
- R. Chandramohanadas *et al.*, *Science* **324**, 794 (2009).
- B. F. Kafsack *et al.*, *Science* **323**, 530 (2009).
- We thank K. Kettleborough, S. Osborne, and colleagues at Medical Research Council (MRC) Technology for provision of compound 1, S. Badrinath, E. Duraisingh, C. Merrick, and B. Coleman for critical reading of the manuscript, and D. Wirth for essential guidance and support. This work was supported by Pediatric Scientist Development Program Fellowship awards K12-HD000850 (J.D.D. and S.D.P.), National Institutes of Health R01 grant R01AI057919 (M.T.D.), a Burroughs Wellcome Fund New Investigator in the Pathogenesis of Infectious Diseases Fellowship (M.T.D.), the MRC UK U117532063 (M.J.B.), a Wellcome Trust Project grant 086550 (C.R.C.), the NIH GM073046 (T.J.W.), and an EU FP7 grant MALSIG 223044 (D.A.B.).

Supporting Online Material

www.sciencemag.org/cgi/content/full/328/5980/910/DC1
Materials and Methods
SOM Text
Figs. S1 to S7
References
Data File S1
10 February 2010; accepted 13 April 2010
10.1126/science.1188191

Small RNA Duplexes Function as Mobile Silencing Signals Between Plant Cells

Patrice Dunoyer,^{1*} Gregory Schott,¹ Christophe Himber,¹ Denise Meyer,¹ Atsushi Takeda,² James C. Carrington,² Olivier Voinnet^{1*}

In the plant RNA interference (RNAi) pathway, 21-nucleotide duplexes of small interfering RNA (siRNA) are processed from longer double-stranded RNA precursors by the RNaseIII Dicer-like 4 (DCL4). Single-stranded siRNAs then guide Argonaute 1 (AGO1) to execute posttranscriptional silencing of complementary target RNAs. RNAi is not cell-autonomous in higher plants, but the nature of the mobile nucleic acid(s) signal remains unknown. Using cell-specific rescue of DCL4 function and cell-specific inhibition of RNAi movement, we genetically establish that exogenous and endogenous siRNAs, as opposed to their precursor molecules, act as mobile silencing signals between plant cells. We further demonstrate physical movement of mechanically delivered, labeled siRNA duplexes that functionally recapitulate transgenic RNAi spread. Cell-to-cell movement is unlikely to involve AGO1-bound siRNA single strands, but instead likely involves siRNA duplexes.

In plants, RNAi spreads over long distances through the vasculature and from cell to cell presumably via plasmodesmata, owing to mobile nucleic acid-based signals (1, 2). Plant

cell-to-cell RNAi movement has defensive and developmental roles: The spread of viral-derived silencing signals probably immunizes surrounding naïve cells (3), whereas movement of silencing signals from endogenous *TRANS-ACTING siRNA (TAS)* (siRNA, small interfering RNA) loci might generate target gene-expression gradients allowing organ polarization (4, 5). Movement of small RNA is also suspected to account for epigenetic reprogramming between pollen nuclei (6). Both viruses and *TAS* loci produce long double-stranded RNA (dsRNA) precursors, subsequently converted into siRNAs by Dicer-like 4

(DCL4), one of four *Arabidopsis* Dicer-like proteins (3, 7). Although DCL4-dependent siRNAs are popular candidates as cell-to-cell RNAi signaling molecules (5, 8), a role for their precursors, including long dsRNA, cannot be excluded. Consequently, the nature of the mobile, silencing nucleic acid(s) remains unknown (9).

In the experimental *SUC:SUL Arabidopsis* system, a long-dsRNA-producing transgene is expressed under the phloem-companion cell-specific promoter, *SUC2* (8). This triggers non-cell-autonomous RNAi of the ubiquitously expressed endogenous *SULFUR (SUL)* mRNA, generating a leaf chlorotic phenotype that expands 10 to 15 cells beyond the vasculature (Fig. 1A). The *SUL* dsRNA is processed into DCL4-dependent 21-nucleotide (nt) and DCL3-dependent 24-nt siRNAs, which are, respectively, mandatory and dispensable for the manifestation of *SUL*-silencing movement (Fig. 1B) (8). To uncover the *SUL*-silencing signal's identity we used the viral silencing suppressor P19, which specifically sequesters 21-base pair (bp) siRNA duplexes, but not their long-dsRNA precursor (10). A *SUC:P19* transgene was introduced into *SUC:SUL* plants. Highly expressing lines displayed no *SUL*-silencing movement, whereas movement remained unaltered in low-expressing lines (Fig. 1A). As expected, both high and low P19 lines accumulated similar amounts of each *SUL* siRNA species, as found in *SUC:SUL* reference plants (Fig. 1, A and B). Immunoprecipitation analyses using transgenic lines expressing different levels of epitope-tagged P19 (*SUC:P19HA*) revealed that *SUL*-silencing suppression was

¹Institut de Biologie Moléculaire des Plantes du CNRS, Université de Strasbourg 12 rue du Général Zimmer, 67084 Strasbourg cedex, France. ²Center for Genome Research and Biocomputing, Oregon State University, Corvallis, OR 97331, USA.

*To whom correspondence should be addressed. E-mail: patrice.dunoyer@ibmp-ulp.u-strasbg.fr (P.D.); olivier.voinnet@ibmp-ulp.u-strasbg.fr (O.V.)

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

The following resources related to this article are available online at www.sciencemag.org (this information is current as of May 13, 2010):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/328/5980/912>

Supporting Online Material can be found at:

<http://www.sciencemag.org/cgi/content/full/science.1185880/DC1>

This article **cites 22 articles**, 7 of which can be accessed for free:

<http://www.sciencemag.org/cgi/content/full/328/5980/912#otherarticles>

This article has been **cited by** 1 articles hosted by HighWire Press; see:

<http://www.sciencemag.org/cgi/content/full/328/5980/912#otherarticles>

This article appears in the following **subject collections**:

Molecular Biology

http://www.sciencemag.org/cgi/collection/molec_biol

correlated with the extent to which 21-nt *SUL* siRNAs were sequestered by P19 (Fig. 1, C to E). Immunostaining in *SUC:P19HA* plants and immunofluorescence analyses carried out in independently generated *SUC:SUL2/SUC:P19HA* lines confirmed that P19HA accumulation is strictly confined within phloem-companion cells (Fig. 1F and fig. S1). The strict cell autonomy of P19 was further confirmed by analysis of independent transgenics in which the protein was driven under a mesophyll-specific promoter (fig. S1). We conclude that suppression of cell-to-cell *SUL*-silencing movement relies on the dose-dependent and cell-autonomous capacity of P19 to specifically sequester DCL4-dependent 21-bp *SUL* siRNAs in silencing-incipient cells. Similar conclusions were drawn using *SUC:SUL* plants expressing companion-cell-specific P21, a distinct silencing suppressor that, like P19, sequesters 21-nt siRNA (fig. S2) (11).

As a second approach to unravel the signal's identity, we investigated cell-type-specific requirements for DCL4 in the *SUL*-silencing movement process. A requirement for DCL4 in incipient, dsRNA-producing companion-cells, but not in recipient cells, would support movement of an siRNA signal (9). A *DCL4* genomic

fragment was cloned under the *SUC2* promoter (*SUC:DCL4*) and transformed into *SUC:SUL* plants carrying a *dcl4-2* null mutation. Without a *DCL4* transgene, *dcl4-2* plants lack 21-nt *SUL* siRNAs and, accordingly, display no *SUL*-silencing movement (Fig. 2, A and B). Lack of *DCL4* results in aberrant processing of endogenous *TAS* precursors into nonfunctional, 22-nt-long siRNAs by DCL2 (Fig. 2B). Consequently, leaves ectopically accumulate high levels of trans-acting siRNA (tasiRNA) targets, including *AUXIN RESPONSE FACTOR 3* (*ARF3*) transcripts, which normally undergo *TAS3* tasiRNA-dependent non-cell-autonomous RNAi over several leaf-primordia cell layers (Fig. 2C) (4, 5). *DCL4* expression was detected in nearly all independently generated *SUC:DCL4/SUC:SUL/dcl4* transgenic lines (fig. S3), and in those lines, 21-nt *SUL* siRNA accumulation was fully rescued (Fig. 2B). Companion-cell-restricted *DCL4* expression also rescued full *SUL*-silencing movement (Fig. 2A). This result suggests that movement of *SUL* long dsRNA, if any, cannot account for the *SUL*-silencing phenotype. Furthermore, 21-nt-long tasiRNA production was also restored in leaves, as was down-regulation of *ARF3* and

other endogenous tasiRNA targets (Fig. 2, B and C). Consistent with this finding, leaf expression of Argonaute 7 (*AGO7*), a limiting factor in the *TAS3* pathway, is restricted to the vasculature and cells near the abaxial surface (12). Collectively, these data indicate that 21-nt siRNAs account for cell-to-cell movement of exogenous, and possibly endogenous, *DCL4*-dependent RNAi.

SUL-silencing requires the specific loading of 21-nt *SUL* siRNA guide strands into AGO1 (8), 1 of 10 *Arabidopsis* AGO proteins that also effects posttranscriptional silencing with endogenous microRNAs (miRNAs) and tasiRNAs (13). Therefore, we investigated whether 21-nt siRNAs move between cells in association with AGO1. AGO1 immunoprecipitation experiments conducted in independent *SUC:SUL* lines expressing the *SUC:P19HA* or *SUC:P19* transgenes showed that approximately half of the 21-nt *SUL* siRNA pool is sequestered away from AGO1 by phloem-specific P19 (Fig. 3A); this sequestration is sufficient to prevent *SUL*-silencing movement (Fig. 1, A and C). Similar observations were made in *SUC:SUL* plants expressing *SUC:P21* or *SUC:P21HA* (fig. S2). As the other half of the companion-cell-derived 21-nt *SUL* siRNA remains loaded

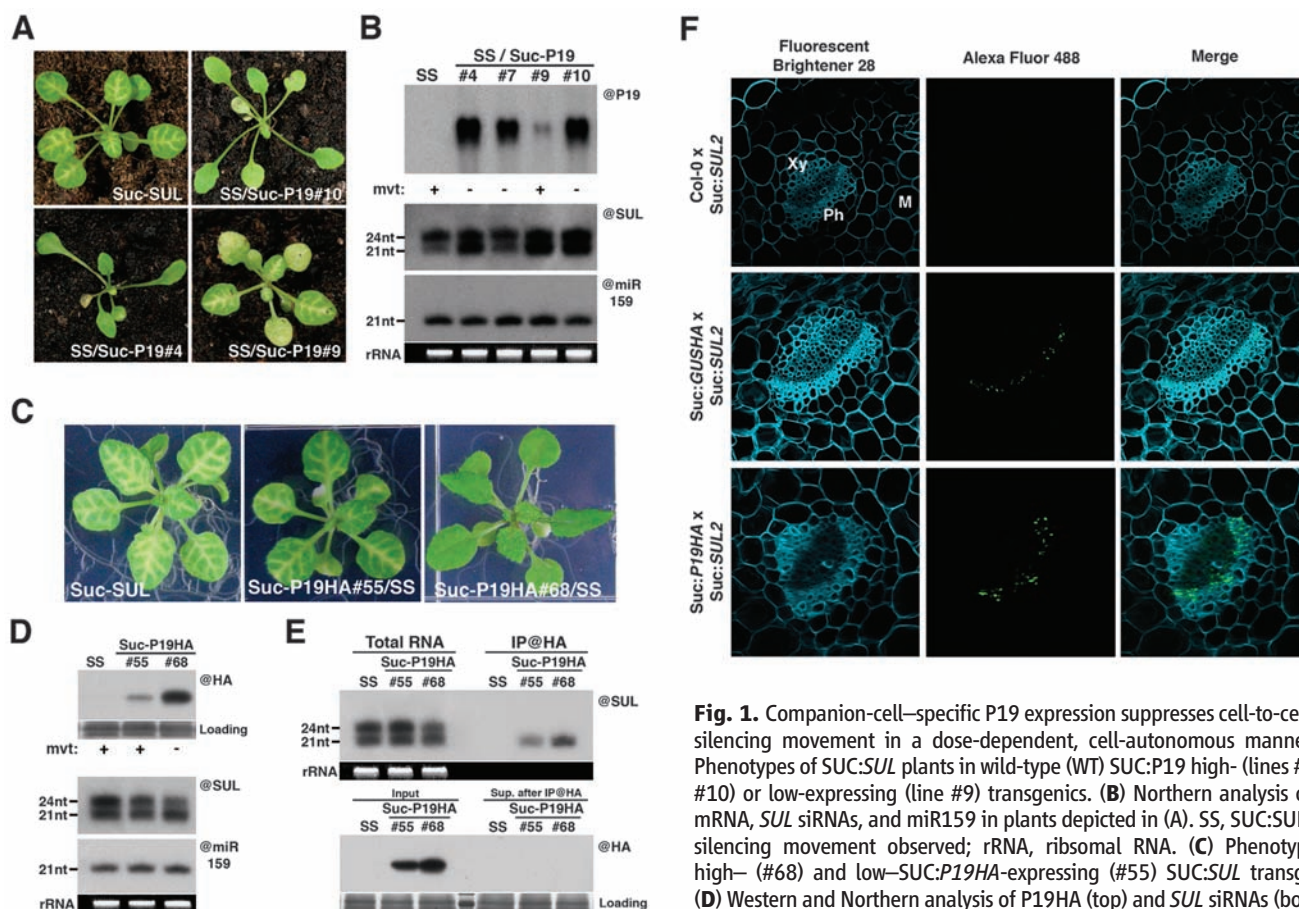


Fig. 1. Companion-cell-specific P19 expression suppresses cell-to-cell *SUL*-silencing movement in a dose-dependent, cell-autonomous manner. (A) Phenotypes of *SUC:SUL* plants in wild-type (WT) *SUC:P19* high- (lines #4 and #10) or low-expressing (line #9) transgenics. (B) Northern analysis of *P19* mRNA, *SUL* siRNAs, and miR159 in plants depicted in (A). SS, *SUC:SUL*; mvt, silencing movement observed; rRNA, ribosomal RNA. (C) Phenotypes of high- (#68) and low-*SUC:P19HA*-expressing (#55) *SUC:SUL* transgenics. (D) Western and Northern analysis of *P19HA* (top) and *SUL* siRNAs (bottom), respectively, in plants depicted in (C). (E) HA epitope-specific immunoprecipitation (IP@HA) of *SUL* siRNAs from *SUC:P19HA* transgenics. Total RNA was extracted from immunoprecipitates (IPs), and low-molecular-weight RNA was subjected to Northern analysis (top). *P19HA* immunoprecipitation was confirmed by protein blot analysis (bottom). (F) *P19HA* immunolocalization (Alexa-Fluor488) in transverse sections of *SUC:P19HA*-expressing *SUC:SUL* transgenics. Plants expressing a *SUC:GUSHA* transgene provide a positive control for companion-cell immunolabeling retention. Cell walls were stained with fluorescent brightener 28. Xy, Xylem; Ph, Phloem; M, Mesophyll.

precipitation in *SUC:SUL* reference plants or in high- and low-*SUC:P19HA*-expressing *SUC:SUL* transgenics. Total RNA was extracted from immunoprecipitates (IPs), and low-molecular-weight RNA was subjected to Northern analysis (top). *P19HA* immunoprecipitation was confirmed by protein blot analysis (bottom). (F) *P19HA* immunolocalization (Alexa-Fluor488) in transverse sections of *SUC:P19HA*-expressing *SUC:SUL* transgenics. Plants expressing a *SUC:GUSHA* transgene provide a positive control for companion-cell immunolabeling retention. Cell walls were stained with fluorescent brightener 28. Xy, Xylem; Ph, Phloem; M, Mesophyll.

into AGO1 (Fig. 3A), it is unlikely that siRNAs move from cell to cell in an AGO1-bound form. Furthermore, AGO1-bound siRNA movement would entail movement of AGO1 itself. To address this issue, a SUC2-driven, functional epitope-

tagged *AGO1* allele (*SUC:FLAG-AGO1*) (14) was transformed into *SUC:SUL* plants carrying either the *ago1-12* or *ago1-27* hypomorphic mutations, both of which prevent *SUL*-silencing movement with minimal or no effect on *SUL* siRNA pro-

duction (Fig. 3, B and D, and fig. S4) (8, 15). Of the five lines inspected, none recovered the *SUL*-silencing movement phenotype (Fig. 3B and fig. S4), despite FLAG-AGO1 being appropriately loaded with the 21-nt *SUL* siRNAs and with endogenous miRNAs (Fig. 3, C and D, and fig. S4). We conclude that AGO1 is required cell-autonomously for the execution of RNAi in incipient and recipient cells. Therefore, DCL4-dependent siRNAs are unlikely to move between cells bound to AGO1.

Next, we performed experiments to directly monitor siRNA movement in plants. We used particle bombardment to mechanically deliver various RNAi trigger molecules into XD216 transgenic seedlings. XD216 contains a constitutively expressed *GREEN FLUORESCENT PROTEIN* (*GFP*) transgene in an *sde1/rdr6* null mutant background (16). This mutation ensures that any silencing events monitored in XD216 are attributable to the bombarded material, as opposed to endogenously amplified and RDR6-dependent secondary silencing events (fig. S5) (16). Bombardment of in vitro transcribed and preannealed *GFP*-derived dsRNA consistently triggered *GFP* silencing: Four days post-bombardment (dpb), it was manifested as foci, 10 to 15 cells in diameter, that did not expand any further (Fig. 4A). The same observation was made by bombarding a mix of three chemically synthesized, 21-bp *GFP* siRNA duplexes (Fig. 4B), but not by bombarding siRNAs or dsRNA with *GFP*-unrelated sequences. Bombarding particles co-coated with a GUS-reporter gene plasmid confirmed that silencing foci resulted from bona fide *GFP*-silencing movement radiating from primarily delivered, individual cells or small groups of cells (Fig. 4, C and D). The *GFP*-silencing–

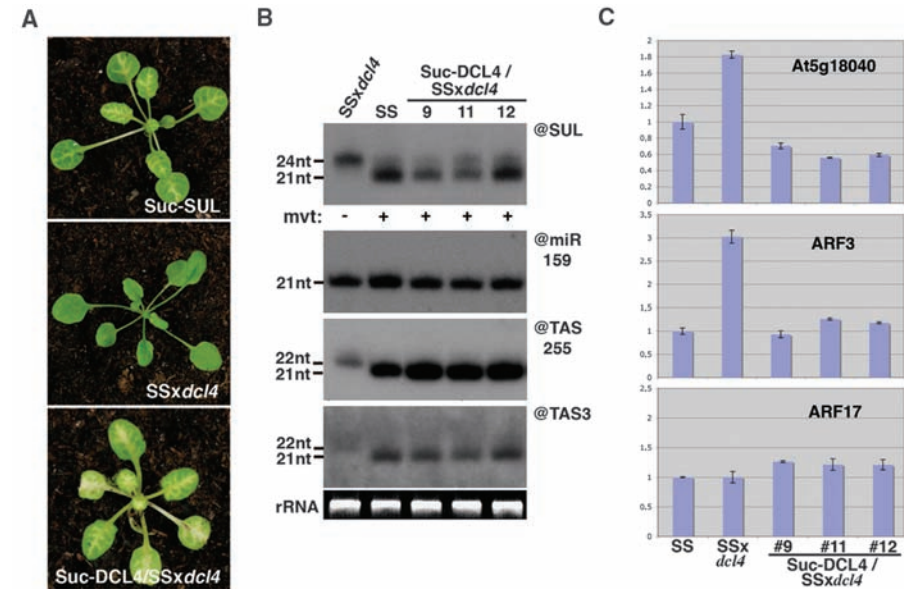
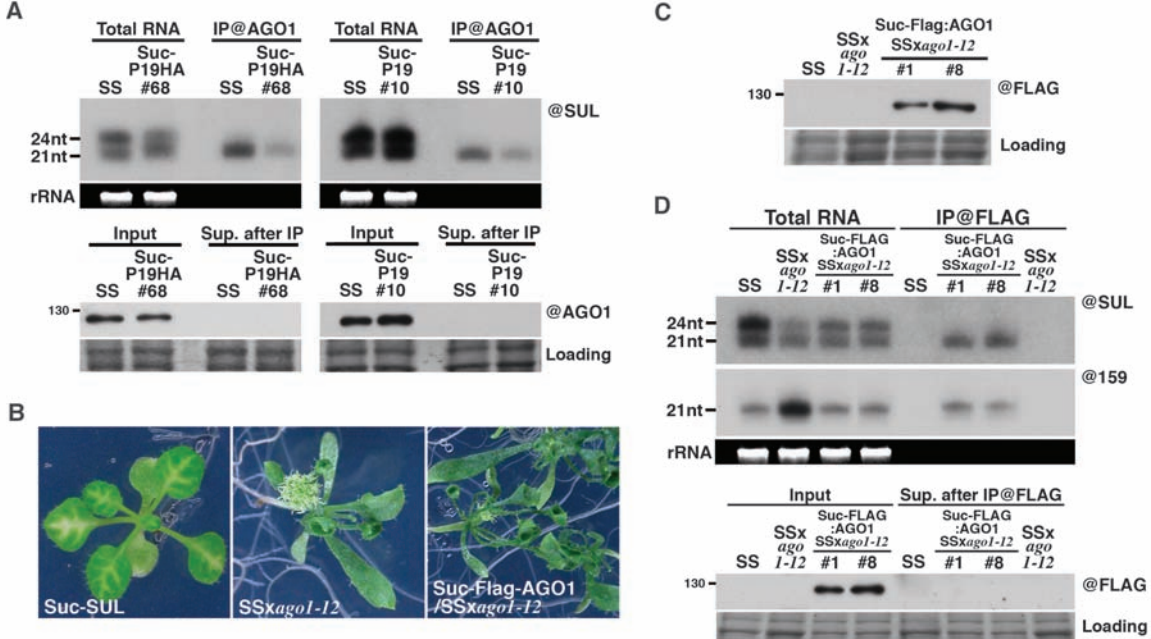


Fig. 2. Companion-cell-specific *DCL4* expression in *SUC:SUL/dcl4* mutants restores 21-nt *SUL* siRNA and tasiRNA accumulation, cell-to-cell *SUL*-silencing movement, and tasiRNA target down-regulation. **(A)** *SUL*-silencing phenotype in WT, *dcl4*, and *SUC:DCL4/dcl4* transgenic plants. **(B)** Northern analysis of *SUL* siRNAs, miR159, TAS1 siRNA255, and TAS3 tasiRNAs in WT, *dcl4*, and independent *SUC:DCL4/dcl4* transgenics (#9, #11, #12). **(C)** Quantitative reverse transcription polymerase chain reaction analysis of endogenous *TAS1 siRNA255* (*At5g18040*) and *TAS3* tasiRNA (*ARF3*) target mRNA accumulation. *ARF17* mRNA (*miR160* target) was used as a nonaffected control. cDNA inputs were normalized to *Actin2* mRNA; expression ratios are relative to levels in the *SUC:SUL* parental line. Data are displayed as averages $\pm$ SD (indicated by error bars) (three replicates).

Fig. 3. Cell-autonomous requirement for AGO1 in *SUL* silencing. **(A)** Northern analysis of total RNA and IP fractions of AGO1-bound small RNA in control, *SUC:P19*-, or *SUC:P19HA*-expressing *SUC:SUL* plants (top). AGO1 immunoprecipitation was confirmed by protein blot analysis (bottom). **(B)** *SUL*-silencing in *SUC:SUL* WT, *ago1-12*, and *SUC:FLAG-AGO1/ago1-12* transgenic plants. **(C)** Western analysis of FLAG-AGO1 in two independent *SUC:FLAG-AGO1/ago1-12* transgenic lines. **(D)** FLAG-specific immunoprecipitation in *SUC:SUL* WT, *ago1-12*, and *SUC:FLAG-AGO1/ago1-12* transgenic plants and Northern analysis of *SUL* siRNA and miR159 accumulation in total RNA or FLAG-AGO1 IP (top). Protein blot analysis confirmed FLAG-AGO1 immunoprecipitation (bottom).



movement process was recapitulated upon bombardment of individual siRNA duplexes (Fig. 4, E and F), but not with single-stranded guide or passenger strands (fig. S5). Moreover, those chemically synthesized siRNA duplexes were faithfully and functionally sorted into cognate AGO effector complexes in bombarded and, presumably, surrounding cells (fig. S5).

To test if *GFP*-silencing foci in XD216 resulted from direct cell-to-cell movement of the bombarded siRNAs, the passenger-strand in the siRNA 6768 duplex was covalently labeled at its 3' end with the fluorophore ALEXA555 (17). One hour pb (hpb), the fluorescence was typically concentrated in single cells, or small groups of cells (Fig. 4G), but by 4 hpb, it had radiated from the initially delivered area into the surrounding cells, a pattern unchanged at 20 hpb and beyond (Fig. 4, H to K). This pattern was unlikely to result from movement of free fluorophore or partially digested siRNAs, as the residual fluorescence was precisely super-

imposed over *GFP*-silencing foci in many areas of bombarded leaves at 4 dpb (Fig. 4, L and M, and fig. S5). The most straightforward interpretation of these results is that labeled siRNA 6768 had moved from the bombarded cells to the adjacent 10 to 15 cells and triggered RNAi. Bombarding the siRNA 6768 duplex with a 3' end-labeled guide strand also triggered movement, but *GFP*-silencing efficacy was reduced, presumably because the ALEXA555 dye prevents optimal loading of guide strands into AGOs. Occasionally, labeled siRNAs reached the vascular system during the cell-to-cell movement process (Fig. 4, N to P). Moreover, singly bombarded phloem cells could clearly transmit siRNAs to adjacent cells within vascular bundles (Fig. 4, Q and R), suggesting that siRNAs may also act as phloem-transported signals for long-distance silencing in *Arabidopsis* (18, 19).

We have provided genetic evidence that siRNAs are necessary for mobile silencing, while the bombardment experiments show that they are

also sufficient for this process. We conclude, therefore, that siRNAs act as silencing signals between plant cells and, possibly, over long distances (19). Although their exact mobile form (that is, protein-bound versus free molecules) awaits further characterization, duplexes, as opposed to single strands, are likely involved. Indeed, movement occurred with preannealed siRNA duplexes containing either a labeled passenger strand or a labeled guide strand. The former is rapidly turned over upon strand separation (20), whereas guide single strands are usually unstable unless loaded into AGOs (21); yet AGO1, the effector of mobile RNAi, functions cell-autonomously (Fig. 3, B to D). Lack of *GFP*-silencing foci upon bombardment of anti-*GFP* single strands also supports movement of siRNAs as duplexes (fig. S5), although single strands might be rapidly degraded in bombarded cells or might fail to incorporate into AGO1 in surrounding cells owing to pre-requisite strand separation. Movement of 21-bp siRNA duplexes between cells and their inhibition by viral P19 agree with previous results obtained with the P19-producing *Cymbidium ringspot virus* (CymRSV): P19 was dispensable for CymRSV accumulation within vascular bundles. However, its lack prevented further invasion of the leaf lamina, which, although virus-free, exhibited nucleotide sequence-specific resistance to CymRSV owing to the onset and cell-to-cell movement of a mobile, virus-induced silencing signal (22). In light of the present results, this signal must be the DCL4-dependent 21-bp siRNAs produced from CymRSV-derived dsRNA.

Of the 21- and 24-bp siRNA species, we studied only the former in this experiment, because AGO1-dependent RNAi generates a measurable silencing movement phenotype (that is, lack of *SUL* or *GFP*). Yet, there is no reason to exclude mobility of DCL3-dependent 24-nt siRNAs, which mediate locus-specific chromatin modifications upon loading into AGO4 (12). Using a micrografting procedure, we have obtained evidence that *Arabidopsis* endogenous siRNAs of all size classes are mobile from cell to cell and over long distances (19). Meanwhile, the experimental system described here now provides a handle to dissect the cell biology of siRNA movement in plants.

References and Notes

1. D. Baulcombe, *Nature* **431**, 356 (2004).
2. O. Voinnet, *FEBS Lett.* **579**, 5858 (2005).
3. A. Deleris *et al.*, *Science* **313**, 68 (2006); published online 1 June 2006 (10.1126/science.1128214).
4. R. Schwab *et al.*, *PLoS ONE* **4**, e5980 (2009).
5. D. H. Chitwood *et al.*, *Genes Dev.* **23**, 549 (2009).
6. R. K. Slotkin *et al.*, *Cell* **136**, 461 (2009).
7. E. Allen, Z. Xie, A. M. Gustafson, J. C. Carrington, *Cell* **121**, 207 (2005).
8. P. Dunoyer, C. Himber, V. Ruiz-Ferrer, A. Alioua, O. Voinnet, *Nat. Genet.* **39**, 848 (2007).
9. P. Dunoyer, O. Voinnet, *Trends Plant Sci.* **14**, 643 (2009).
10. J. M. Vargason, G. Szittyta, J. Burguán, T. M. Hall, *Cell* **115**, 799 (2003).
11. Z. Mérai *et al.*, *J. Virol.* **80**, 5747 (2006).
12. T. A. Montgomery *et al.*, *Cell* **133**, 128 (2008).
13. H. Vaucheret, *Trends Plant Sci.* **13**, 350 (2008).

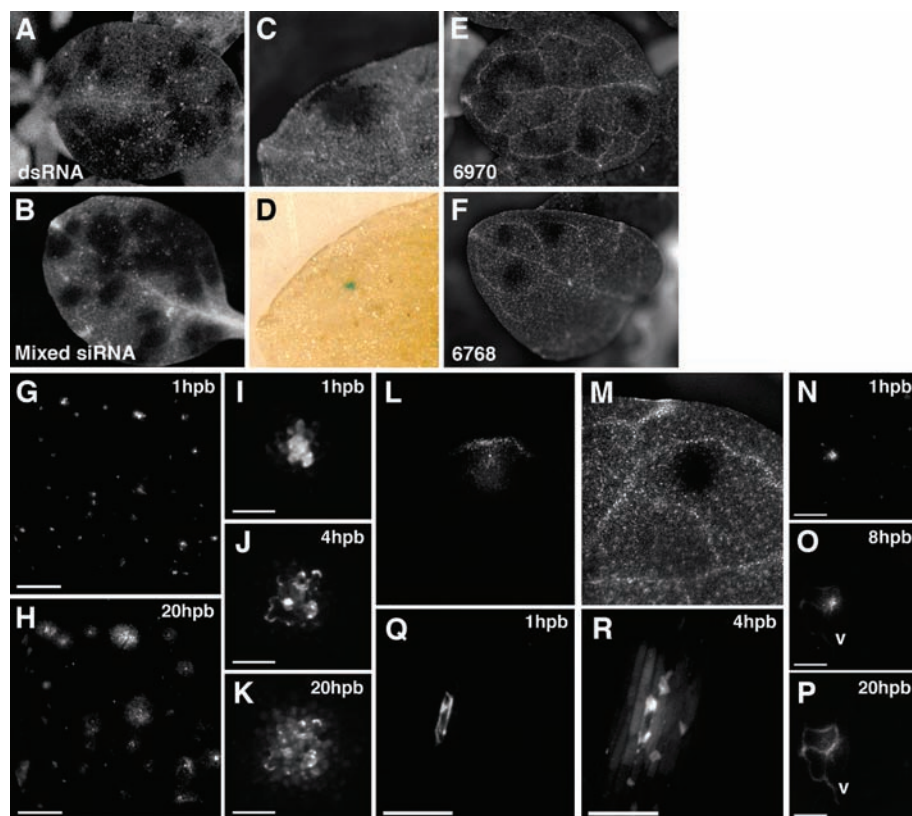


Fig. 4. Bombarded siRNAs trigger GFP silencing and move from cell to cell. (A and B) GFP-silencing foci on XD216 leaves triggered by *GFP*-derived dsRNA or a mix of three independent, chemically synthesized siRNAs duplexes. (C and D) An XD216 leaf cobombarded with *GFP*-targeting siRNA and *GUS*-expressing plasmid. *GFP* silencing is visible at 4 dpb (C). *GUS* staining was done at 4 dpb (D). (E and F) *GFP*-silencing foci on XD216 triggered by single *GFP*-targeting siRNA duplex 6768 or 6970. (G and H) *Arabidopsis* Col-0 leaf bombarded with ALEXA555-labeled siRNA 6768 at 1 hpb (G) and 20 hpb (H). (I to K) Same as in (G) and (H) at 1 hpb (I), 4 hpb (J), and 20 hpb (K) on turnip leaves. (L and M) Coincidence of ALEXA555-labeled siRNA 6768 (L) and *GFP* silencing (M) on XD216 at 4 dpb. (N and P) Bombarded Col-0 leaf at 1 hpb (N), 8 hpb (O), 20 hpb (P) showing that ALEXA555-labeled siRNA 6768 reaches the vascular system (v). (Q and R) Movement of ALEXA555-labeled siRNA 6768 within the vascular bundle at 1 hpb (Q) and 4 hpb (R). Scale bars: 500 μ m for (G), (H), (N), (O), and (P); 100 μ m for (I), (J), and (K); 200 μ m for (Q) and (R).

14. N. Baumberger, D. C. Baulcombe, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 11928 (2005).
15. P. Brodersen *et al.*, *Science* **320**, 1185 (2008); published online 15 May 2008 (10.1126/science.1159151).
16. C. Himber, P. Dunoyer, G. Moissiard, C. Ritzenthaler, O. Voinnet, *EMBO J.* **22**, 4523 (2003).
17. S. Y. Berezina, L. Supekova, F. Supek, P. G. Schultz, A. A. Deniz, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 7682 (2006).
18. O. Voinnet, P. Vain, S. Angell, D. C. Baulcombe, *Cell* **95**, 177 (1998).
19. D. S. Schwarz *et al.*, *Cell* **115**, 199 (2003).
20. D. S. Schwarz, G. Hutvagner, B. Haley, P. D. Zamore, *Mol. Cell* **10**, 537 (2002).
21. Z. Havelda, C. Hornyik, A. Crescenzi, J. Burgyn, *J. Virol.* **77**, 6082 (2003).
22. P. Dunoyer *et al.*, *EMBO J.*, published online 22 April 2010 (10.1038/emboj.2010.65).
23. Research in O.V.'s laboratory is funded by a prize from the Bettencourt Foundation for Life Science Research and a starting grant from the European Research Council "Frontiers of RNAi" ERC 210890. P.D. and G.S. are supported by a research grant from Agence National pour la Recherche (ANR-08-JCJC-0063-01), J.C.C. is supported by grants from the NIH (AI43288) and the NSF (MCB-0618433), and A.T. is supported by a fellowship from the Japanese Society for the Promotion of Science. We thank J. Mutterer for help with image

acquisition, R. Wagner's team for plant care, and members of O.V.'s laboratory for critical reading of the manuscript.

Supporting Online Material

www.sciencemag.org/cgi/content/full/science.1185880/DC1
SOM Text
Figs. S1 to S5
References

14 December 2009; accepted 28 January 2010
Published online 22 April 2010;
10.1126/science.1185880
Include this information when citing this paper.

Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation

Assaf Zemach, Ivy E. McDaniel, Pedro Silva, Daniel Zilberman*

Eukaryotic cytosine methylation represses transcription but also occurs in the bodies of active genes, and the extent of methylation biology conservation is unclear. We quantified DNA methylation in 17 eukaryotic genomes and found that gene body methylation is conserved between plants and animals, whereas selective methylation of transposons is not. We show that methylation of plant transposons in the CHG context extends to green algae and that exclusion of histone H2A.Z from methylated DNA is conserved between plants and animals, and we present evidence for RNA-directed DNA methylation of fungal genes. Our data demonstrate that extant DNA methylation systems are mosaics of conserved and derived features, and indicate that gene body methylation is an ancient property of eukaryotic genomes.

Our knowledge about cytosine methylation has been derived mainly from four species: humans, mouse, *Arabidopsis thaliana*, and *Neurospora crassa* (1). Transposons and repeats are virtually uniformly methylated in all four, which suggests that transposon defense is an ancient function of methylation (1, 2). The expression of imprinted genes is regulated by DNA methylation in *A. thaliana* and mammals (3), and in both cases the bodies of active genes are methylated (4–6). However, these analogies do not necessarily imply conservation. Regulation of imprinting has evolved independently in mammals and flowering plants (3). Many transposons are not methylated in the tunicate *Ciona intestinalis* (7), and the transposon-rich silk moth (*Bombyx mori*) genome contains low levels of 5-methylcytosine (8), putting into question the conservation of transposon methylation (9). To understand how eukaryotic DNA methylation has evolved, we quantified DNA methylation by deep bisulfite sequencing in the genomes of five plants, seven animals, and five fungi (figs. S1 to S6 and tables S1 to S3). We also profiled transcription in these species by deep sequencing of cDNA (10). The results reveal a complex evolutionary history of the DNA meth-

ylation pathway and allow us to reconstruct a plausible ancestral state.

Rice (*Oryza sativa*) is a model monocot that contains Dnmt1 (CG methyltransferase), Dnmt3 [de novo methyltransferase responsible for plant CHH methylation (H = A, C, or T)], and CMT (plant-specific CHG methyltransferase) orthologs (1) (figs. S1 to S5 and table S4). Our data show that CG methylation is lowest from 100 base pairs (bp) upstream of the transcriptional start site (TSS) of rice genes to 500 bp within the transcript, plateauing around 1.5 kb after the TSS (Fig. 1A). Non-CG methylation is essentially absent from genes, whereas methylation in all contexts is abundant in transposable elements (TEs; Fig. 1, A and B), with short TEs particularly enriched in CHH methylation (fig. S7). Consistent with functioning to repress transcription, gene expression varies inversely with methylation of the TSS-proximal region (Fig. 1C). The same pattern is present at the 3' end of genes, which suggests that lack of methylation around the transcription termination site is also important for gene expression (Fig. 1C). Within the gene body plateau, rice methylation exhibits a parabolic relationship with transcription: Modestly expressed genes are most likely to be methylated, whereas genes at either transcriptional extreme are least likely to be methylated (Fig. 1C and fig. S8). In all salient features, rice methylation patterns closely resemble those of *A. thaliana* (5, 6).

We examined methylation in two early-diverging land plants: *Selaginella moellendorffii*

and *Physcomitrella patens* (fig. S1) (11), both possessing Dnmt1, Dnmt3, and CMT orthologs (figs. S1 to S5). *S. moellendorffii* TEs are methylated in the CG, CHG, and CHH contexts, but unlike the angiosperms, genes have little methylation, and there is essentially no methylation around the TSS regardless of transcription (Fig. 1, D to F, and table S1). *P. patens* methylation patterns closely resemble those of *S. moellendorffii* (table S1 and fig. S9) (10). DNA methylation in these plants is strictly segregated away from genes.

We analyzed two green algae, *Chlorella* sp. NC64A and *Volvox carteri* (figs. S1 to S3). *Chlorella* has the greatest amount of CG methylation of any organism analyzed here (Fig. 1, G to I, and table S1): Genes are methylated virtually without exception (fig. S10), with a sharp drop of methylation at the promoter (Fig. 1G). Promoter methylation correlates negatively with gene expression (Fig. 1I), which suggests that TSS-proximal methylation represses transcription, as it does in land plants. Although CMT genes have thus far been described only in land plants (1), *Chlorella* contains a CMT homolog (figs. S1 to S3) and a substantial amount of CHG methylation (table S1), which is concentrated in repetitive elements and excluded from genes (Fig. 1, G to I). The *V. carteri* genome is much less methylated exclusively in the CG context, with a similar relationship between promoter methylation and transcription, and displays preferential methylation of TEs (table S1 and fig. S9). Methylation of both gene bodies and TEs thus appears to be an ancient property of plants.

The genome of the puffer fish *Tetraodon nigroviridis* is virtually devoid of transposable elements (12), yet we find heavy methylation exclusively at CG sites, the vast majority of which is found in genes, although TEs do show particularly dense methylation (Fig. 2, A and B, and table S1). Genes exhibit a prominent dip in methylation just upstream of the TSS, and reduced promoter methylation correlates with enhanced expression (Fig. 2A). Gene body methylation exhibits the same relationship with transcription that exists in *A. thaliana* and rice—a roughly parabolic curve with the most methylated genes around the 70th transcription percentile (fig. S8).

The similarities between plants and vertebrates prompted us to ask whether aspects of the

Department of Plant and Microbial Biology, 211 Koshland Hall, University of California, Berkeley, CA 94720, USA.

*To whom correspondence should be addressed. E-mail: danielz@berkeley.edu

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

The following resources related to this article are available online at www.sciencemag.org (this information is current as of May 13, 2010):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/328/5980/916>

Supporting Online Material can be found at:

<http://www.sciencemag.org/cgi/content/full/science.1186366/DC1>

This article **cites 15 articles**, 3 of which can be accessed for free:

<http://www.sciencemag.org/cgi/content/full/328/5980/916#otherarticles>

This article has been **cited by 1** articles hosted by HighWire Press; see:

<http://www.sciencemag.org/cgi/content/full/328/5980/916#otherarticles>

This article appears in the following **subject collections**:

Molecular Biology

http://www.sciencemag.org/cgi/collection/molec_biol

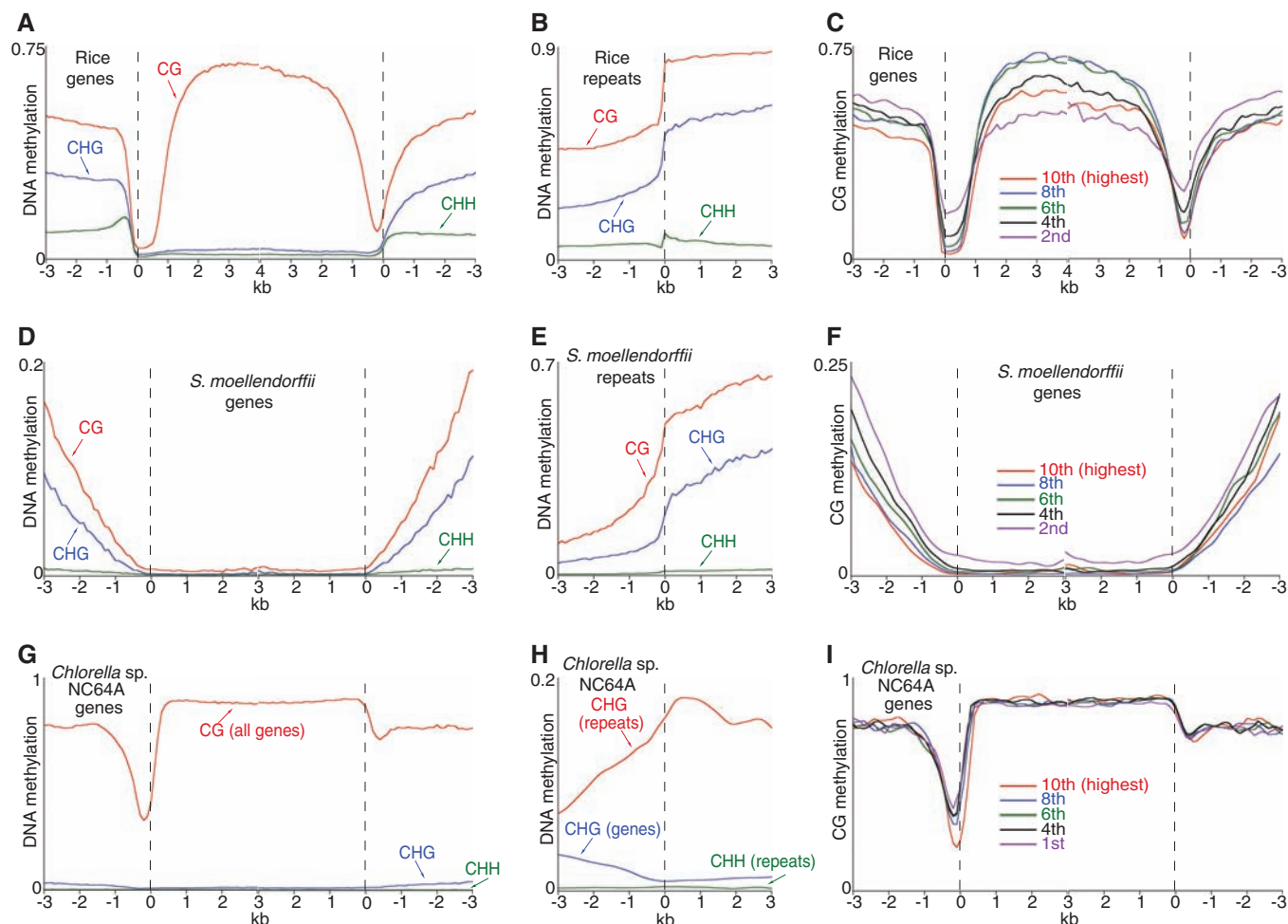
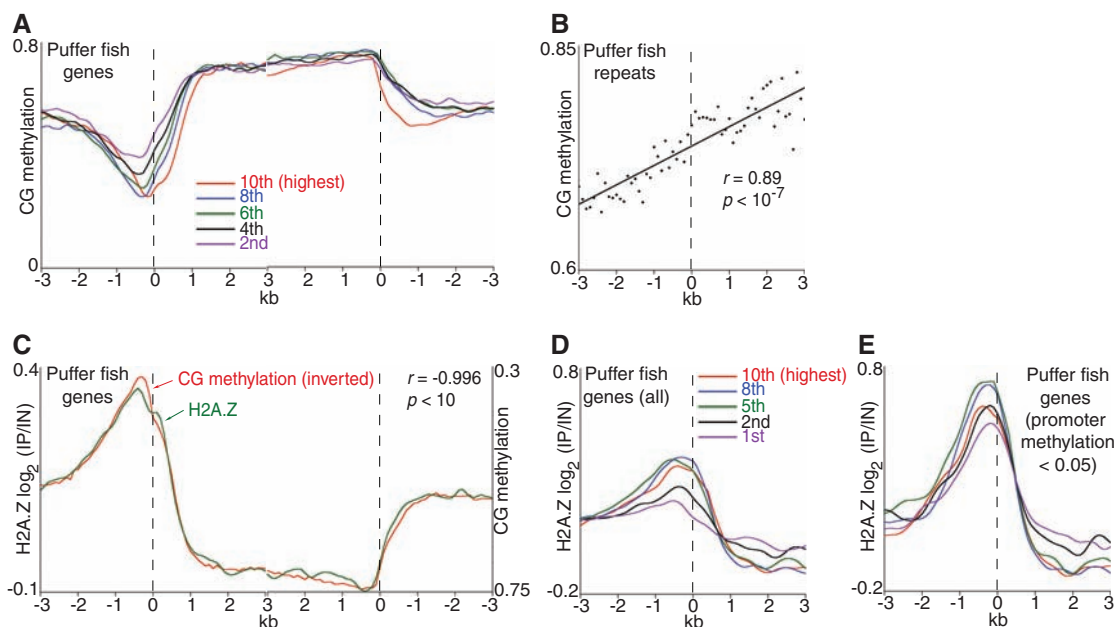


Fig. 1. DNA methylation in gene bodies and repeats of plants: (A to C) rice, (D to F) *S. moellendorffii*, and (G to I) *Chlorella* sp. NC64A. Genes or repeats were aligned at the 5' end (left dashed line) or the 3' end (right dashed line), and average methylation levels for each 100-bp interval are plotted. In (C), (F), and (I), genes were grouped into deciles by transcription, with five of the deciles shown for clarity. Only 5' alignments are shown in (B), (E), and (H).

Fig. 2. Puffer fish DNA methylation and H2A.Z are anti-correlated. (A to E) Genes or repeats were aligned as in Fig. 1, and average methylation levels [(A) to (C)] or H2A.Z enrichment [(C) to (E)] for each 100-bp interval are plotted. In (A), (D), and (E), genes were grouped into deciles by transcription. Only 5' alignments are shown in (B), (D), and (E). In (C) to (E), IP, immunoprecipitated; IN, input control.



interaction between DNA methylation and chromatin are conserved. We recently showed that the histone variant H2A.Z exhibits a strong antagonism with DNA methylation in *A. thaliana* (13). Using chromatin immunoprecipitation coupled with deep sequencing, we found that puffer fish H2A.Z peaks in the promoter and is depleted in gene bodies, with a pattern that is almost precisely opposite to that of DNA methylation (Fig. 2C). H2A.Z is enriched in unmethylated promoters regardless of transcription (Fig. 2, D and E) and is quantitatively depleted in methylated gene bodies (fig. S11). Within unmethylated promoters, H2A.Z exhibits a parabolic distribution relative to transcription; H2A.Z is most enriched around the 50th expression percentile (Fig. 2E and fig. S11). Within gene bodies, H2A.Z enrichment varies inversely with transcription (fig. S11). These features are essentially the same as in *A. thaliana* (13), indicating that the interactions among H2A.Z, DNA methylation, and transcription are conserved between angiosperms and vertebrates.

We examined six invertebrate species belonging to three major groups that diverged 900 million to 1 billion years ago (fig. S1) (11). Flour beetle (*Tribolium castaneum*) adults and *Drosophila melanogaster* embryos aged 0 to 3 hours do not have detectable DNA methylation of the nuclear genome (table S1 and fig. S12) (10). The general methylation features of the other four species [*C. intestinalis*, honey bee (*Apis mellifera*), silk moth, and anemone (*Nematostella vectensis*)] are quite similar (Fig. 3 and table S1). Gene bodies are methylated in roughly the same pattern as plant and fish genes, with highest methylation of moderately expressed genes but no correlation between TSS methylation and transcription (Fig. 3 and fig. S8). TEs are hypomethylated, with methylation rising linearly with distance from the TE (Fig. 3 and fig. S10; we did not perform linear correlation analysis in honey bees because of low TE content). There is thus little evidence that methylation inhibits transcription or silences TEs in invertebrates. Our data indicate that gene body methylation is basal, predating the divergence of plants and animals around 1.6 billion years ago (fig. S1), whereas the antitransposon function probably evolved independently in the vertebrate and plant lineages.

We analyzed five species belonging to the three major fungal groups. *Phycomyces blakesleeanus* diverged from the Dikarya (ascomycetes and basidiomycetes) more than 1 billion years ago (fig. S1) (11) and has two methyltransferases: a Dnmt1-like and a Dim-2-like (figs. S1 to S3). The presence of Dim-2, previously described only in ascomycetes (1), demonstrates that this is an ancient methyltransferase family. We found a substantial amount of CG methylation and a small amount of non-CG methylation (table S1), both concentrated in transcriptionally silent, repetitive loci (Fig. 4, A and B, and fig. S13); methylation in this fungus appears to be used to silence TEs and other repeats, whereas active

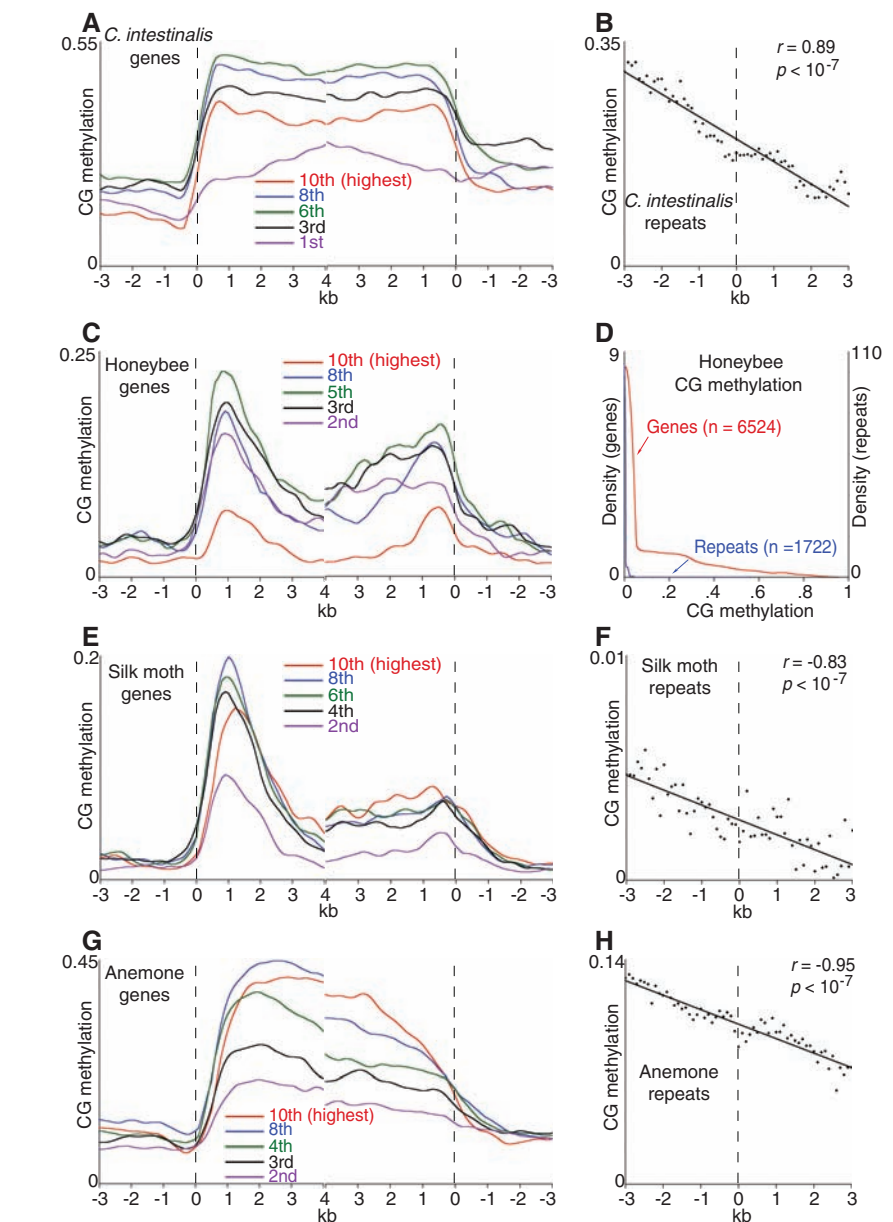


Fig. 3. Invertebrate gene bodies, but not TEs, are preferentially methylated: (A and B) *C. intestinalis*, (C and D) honey bee, (E and F) silk moth, (G and H) anemone. Genes or repeats were aligned as in Fig. 1. In (A), (C), (E), and (G), genes were grouped into deciles by transcription. Only 5' alignments are shown in (B), (F), and (H). (D) Kernel density plot, which has the effect of tracing the frequency distribution, of the average CG methylation within honey bee genes (red trace) and repeats (blue trace).

genes are unmethylated. The three basidiomycete species we examined (*Coprinopsis cinerea*, *Laccaria bicolor*, and *Postia placenta*) show very similar methylation patterns (Fig. 4, C and D, table S1, and figs. S13 and S14), indicating that this is an ancient state that was present in the last common ancestor of most, if not all, modern fungi.

Uncinocarpus reesii is an ascomycete that is distantly related to *Neurospora crassa* (fig. S1). We found that *U. reesii* methylation is generally similar to that of *N. crassa* (14): Silent repeated loci are heavily methylated in all sequence contexts, with little methylation else-

where (Fig. 4, E to G). There is no preference for CG sites, consistent with lack of Dnmt1 (Fig. 4, E to G, and table S1). In fact, *U. reesii* repeats have very little CG methylation (Fig. 4, E to G), because the repeats contain only 2% of the expected frequency of CG sites, with a corresponding increase of CA and TG dinucleotides (fig. S15), the expected products of C/G to T/A deamination. Our observations are consistent with repeat-induced point mutation (RIP), a DNA methylation-linked process that causes mutation of repeated fungal sequences (15). *N. crassa* RIP was suggested to be an evolutionary hindrance because the process limits evolution

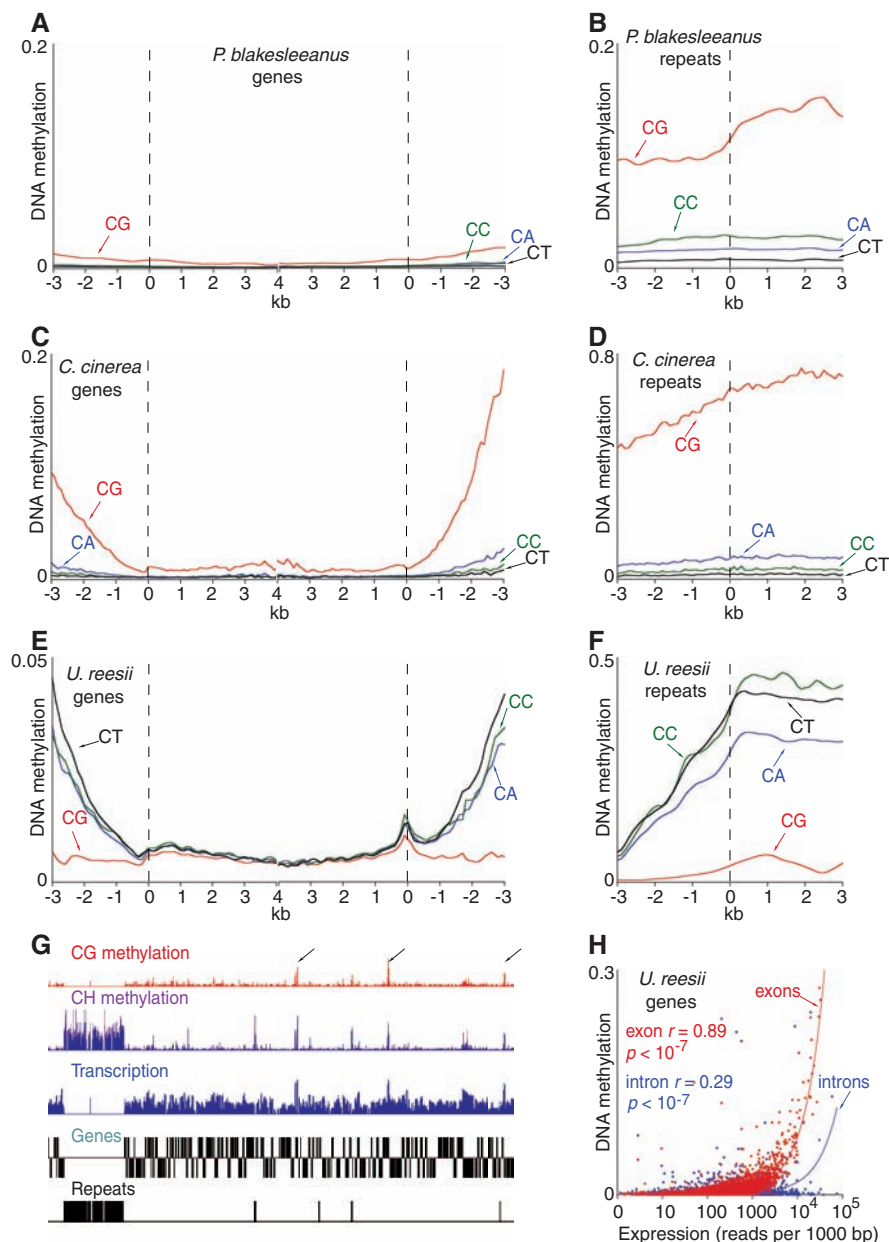


Fig. 4. DNA methylation patterns in fungi: (A and B) *P. blakesleeenanus*, (C and D) *C. cinerea*, (E and F) *U. reesii*. Genes or repeats were aligned as in Fig. 1. Only 5' alignments are shown in (B), (D), and (F). (G) A representative snapshot of *U. reesii* chromosome 1 (positions 3,931,000 to 4,424,000). Arrows highlight CG methylation peaks corresponding to highly transcribed genes. (H) Scatterplot of *U. reesii* DNA methylation in exons (red trace) and introns (blue trace) versus genic RNA levels.

by gene duplication (15), but the presence of a severe version of RIP in an organism separated from *N. crassa* by more than 600 million years would indicate that RIP is a stable evolutionary strategy.

Unlike all other examined fungi, *U. reesii* exhibits methylation of active genes, which is not accompanied by RIP, so CG methylation is abundant (Fig. 4, E, G, and H). There is a direct linear relationship between RNA levels and methylation of exons, whereas introns tend to be unmethylated (Fig. 4H and fig. S16). These features strongly argue that genic DNA methylation is directed by spliced mRNA. It is remarkable that

U. reesii, which possesses none of the methyltransferases of plants, evolved to methylate both TEs and active genes, and, like plants, can distinguish between them.

Our data allow us to undertake a reconstruction of the DNA methylation machinery and functionality of the last common ancestor of plants, animals, and fungi. This organism possessed Dnmt1, Dnmt3, and probably a CMT/Dim-2 enzyme; structural, functional, and phylogenetic data strongly argue that CMT and Dim-2 are monophyletic (figs. S2, S3, and S17) (10). The methylation landscape—transcription-influenced CG methylation of gene bodies and

extensive CG and non-CG methylation of TEs—resembled that of extant angiosperms. H2A.Z was excluded from methylated DNA. The ancestral system then underwent varying degrees of change. A clue to this plasticity is suggested by the observation that all examined land plants and vertebrates retain extensive DNA methylation (1). These groups share two characteristics: They use methylation to repress TEs, and they reproduce primarily by sexual outcrossing. The degree of sexual outcrossing correlates with TE aggressiveness, because outcrossing partially breaks the link between host and TE fitness (16), so land plants and vertebrates are under strong selection to maintain effective TE suppression. Early unicellular animals and fungi that primarily reproduced asexually were more at liberty to lose TE methylation. In contrast, gene body methylation is likely a double-edged sword: It is mutagenic (1), so whatever benefits are derived by methylating exons come at a price (10). Loss of this pathway would thus be more common, and in fact did occur early in fungal evolution and later in several plant and animal lineages.

Note added in proof: Similar conclusions regarding the conservation of gene body methylation have been reached by Feng *et al.* (17).

References and Notes

1. M. G. Goll, T. H. Bestor, *Annu. Rev. Biochem.* **74**, 481 (2005).
2. I. R. Henderson, S. E. Jacobsen, *Nature* **447**, 418 (2007).
3. R. Feil, F. Berger, *Trends Genet.* **23**, 192 (2007).
4. R. Lister *et al.*, *Nature* **462**, 315 (2009).
5. X. Zhang *et al.*, *Cell* **126**, 1189 (2006).
6. D. Zilberman, M. Gehring, R. K. Tran, T. Ballinger, S. Henikoff, *Nat. Genet.* **39**, 61 (2006).
7. M. W. Simmen *et al.*, *Science* **283**, 1164 (1999).
8. C. V. Patel, K. P. Gopinathan, *Anal. Biochem.* **164**, 164 (1987).
9. M. M. Suzuki, A. Bird, *Nat. Rev. Genet.* **9**, 465 (2008).
10. See supporting material on Science Online.
11. S. B. Hedges, J. Dudley, S. Kumar, *Bioinformatics* **22**, 2971 (2006).
12. O. Jaillon *et al.*, *Nature* **431**, 946 (2004).
13. D. Zilberman, D. Coleman-Derr, T. Ballinger, S. Henikoff, *Nature* **456**, 125 (2008).
14. Z. A. Lewis *et al.*, *Genome Res.* **19**, 427 (2009).
15. J. E. Galagan, E. U. Selker, *Trends Genet.* **20**, 417 (2004).
16. T. H. Bestor, *Genetica* **107**, 289 (1999).
17. S. Feng *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 10173 (2010).
18. Acknowledgments are included in supporting material on Science Online. Sequencing data are deposited in Gene Expression Omnibus with accession number GSE19824.

Supporting Online Material

www.sciencemag.org/cgi/content/full/science.1186366/DC1
Materials and Methods
SOM Text
Figs. S1 to S17
Tables S1 to S4
References

23 December 2009; accepted 5 March 2010
Published online 15 April 2010;
10.1126/science.1186366
Include this information when citing this paper.

PROTEOMICS: MASS SPECTROMETRY

PROTEIN SOLUTION CONCENTRATION

The miVac sample concentrator provides an efficient and cost-effective alternative to traditional membrane centrifugation techniques for the concentration of protein solutions prior to separation through size-exclusion chromatography or electrophoresis or to analysis with X-ray crystallography. The membrane centrifugation process often results in loss of valuable sample because of protein binding to the membrane. The miVac sample concentrator ensures complete sample recovery even when taking proteins to very high concentrations. The miVac also offers the benefit of reduced spending on consumables. The instrument provides high performance, precise temperature control, and ease of use within a small benchtop footprint.

Genevac

For info +44-1473-240000 | www.genevac.co.uk

SINGLE QUADRUPOLE GC/MS

The Thermo Scientific ISQ single quadrupole gas chromatography/mass spectrometry (GC/MS) system features a unique, nonventing, full-source removal capability. Featuring the new ExtractaBrite ion source, the ISQ system offers simplified operation and maximum uptime for uninterrupted productivity across routine GC/MS applications, including forensics, toxicology, food safety, and environmental. This robust and rugged system provides high throughput sample analyses and 24/7 operation. The ExtractaBrite removable ion source is designed to provide maximum uptime and productivity by extending the length of time between required source maintenance. For today's fast-paced gas chromatography methods, advanced electronics allow the ISQ to acquire and write data to disc at accelerated rates for high-speed data acquisition in real-world applications.

Thermo Fisher Scientific

For info: 800-532-4752 | www.thermo.com/getready

PEPTIDE SYNTHESIZER

The Syro Wave is a microwave and parallel peptide synthesizer that combines the proven performance of the established MultiSynTech robotic synthesizer with Biotage microwave technology. It is the only system on the market to offer both microwave and parallel peptide synthesis capabilities. Prior to this combination, peptide synthesis labs have had to invest in both standalone parallel systems for productivity and cost efficiency, and standalone microwave systems for difficult or longer peptides. The new system increases productivity, yield, and purity while cutting costs and saving time.

Biotage

For info: +46-18-56-59-00 | www.biotage.com

TIME-OF-FLIGHT SPECTROMETER

The BenchTOF-dx mass spectrometer is a compact reflectron time-of-flight (TOF) instrument designed to generate classical electron impact ionization spectra for gas chromatography (GC) and comprehensive GC applications. The BenchTOF-dx enhances the benefits of standard TOF spectrometers for better handling of challenging compounds. The instrument's capability to generate classical spectra enables laboratories with proprietary spectral libraries to harness the performance of TOF mass spectrometry while maintaining the use of their current library system. The instrument incorporates groundbreaking new ion optics and electronics to deliver full spectral information with sensitivity normally associated only with the best quadrupoles running in selected ion monitoring mode.

ALMSCO

For info: +44-(0)-1443-233920 | www.almsco.com

HUMAN CYTOKINE ARRAY

The Quantibody Human Cytokine Array 4000 can detect quantitatively the expression of 200 different proteins in a single experiment. The kit consists of five quantitative, multiplex enzyme-linked immunosorbent assay-based slides: The Quantibody Human Inflammation Array 3 detects 40 inflammation markers. The Quantibody Human Chemokine Array 1 detects 40 chemokines. The Quantibody Human Growth Factor Array 1 detects 40 growth factors. The Quantibody Human Receptor Array 1 detects 40 soluble cytokine receptors. The Quantibody Human Cytokine Array 4 detects 40 cytokines and related proteins. These glass-slide antibody arrays feature standard concentration curves run alongside unknown samples to give quantitative results with picogram-per-milliliter sensitivity.

RayBiotech

For info: 888-494-8555 | www.raybiotech.com

Electronically submit your new product description or product literature information! Go to www.sciencemag.org/products/newproducts.dtl for more information. Newly offered instrumentation, apparatus, and laboratory materials of interest to researchers in all disciplines in academic, industrial, and governmental organizations are featured in this space. Emphasis is given to purpose, chief characteristics, and availability of products and materials. Endorsement by *Science* or AAAS of any products or materials mentioned is not implied. Additional information may be obtained from the manufacturer or supplier.